Temas de matemáticas

Estadística descriptiva

Luis Rincón

las prensas de ciencias

Luis Rincón

Estadística descriptiva

Facultad de Ciencias, UNAM
2017

Esta obra contó con el financimiento del proyecto PAPIME PE_101216

Estadística descriptiva
1ª edición, 28 de septiembre de 2017

Facultad de Ciencias
Ciudad Universitaria, Delegación Coyoacán.
C.P. 04510. Ciudad de México
editoriales@ciencias.unam.mx

ISBN: 978-607-02-9724-3
Diseño de portada: Laura Uribe

Impreso y hecho en México.

Prólogo

Este pequeño libro contiene temas sobre algunos conceptos de la estadística descriptiva. Los temas que aquí se presentan aparecen en la primera parte del temario oficial de un primer curso de estadística matemática que se ofrece a estudiantes de las carreras de actuaría y matemáticas aplicadas en la Facultad de Ciencias de la UNAM. Esta asignatura es obligatoria para las carreras mencionadas y optativa para las otras carreras que se imparten en la Facultad de Ciencias.

Dada la extrema utilidad y ubicuidad de los métodos estadísticos, muchas otras carreras que se imparten en las distintas escuelas y facultades de la UNAM contemplan por lo menos un curso de estadística, a un nivel adecuado para cada carrera o programa de estudios, tanto de nivel licenciatura como de posgrado. Y sin duda todas estas asignaturas incluyen en sus temarios oficiales algunos aspectos de la estadística descriptiva como una introducción o motivación para el estudio y aplicación de métodos más avanzados de la estadística. Esta es la razón por la que el presente trabajo está dirigido a un amplio público de estudiantes y profesionistas, quienes en sus muy diversas áreas del conocimiento desean conocer alguna técnica o procedimiento para describir un conjunto de datos o resultados obtenidos de algún estudio estadístico, o bien para entender con un poco más de conocimientos las distintas descripciones estadísticas que de manera cotidiana recibimos todos como información.

El lector encontrará aquí diversos temas tradicionales de la estadística descriptiva: clasificación de variables, escalas de medición, medidas de localización y de dispersión, y descripciones gráficas para un conjunto de datos. Se ha buscado que en el texto aparezca el mayor número de gráficas posibles y se ha incorporado un conjunto de ejercicios para que puedan ser desarrollados en el salón de clase o bien para dejarlos de tarea durante el desarrollo de un curso. Puesto que con este trabajo se busca llegar a un público lo más amplio posible, se ha hecho un esfuerzo para que la presentación de los temas y las explicaciones sean de un nivel matemático mínimo.

En donde ha sido posible y adecuado, se ha ilustrado una manera breve de usar el paquete estadístico R para obtener los resultados de algunos cálculos en una computadora. Sin embargo, como nuestro objetivo principal en este trabajo es el de explicar los conceptos de la estadística descriptiva, no se profundiza ni se hace énfasis en el uso de esta herramienta computacional. Así, se han incorporado algunos usos del paquete R para beneficio de aquellos lectores con algunos conocimientos o deseos de conocer esta herramienta computacional, sin que sea absolutamente necesario su uso para el entendimiento de los temas estudiados.

Agradezco a la DGAPA UNAM por el apoyo otorgado a través del proyecto PAPIME PE101216, mediante el cual pudo ser posible la edición de este trabajo. Asimismo, agradezco muy sinceramente a los árbitros de este trabajo por sus valiosos comentarios, y a Patricia Magaña y a Mercedes Perelló por su excelente labor editorial.

Luis Rincón
Agosto 2017
Ciudad Universitaria · UNAM

Contenido

1. Conceptos elementales **1**
Población y muestra 1
Variables y datos 4
Clasificación de variables 5
Escalas de medición 7
Agrupamiento de valores 11
Sobre los datos 14
¿Qué es la estadística descriptiva? 15
▪ Ejercicios 16

2. Descripciones numéricas **23**
Medidas de localización 24
Medidas de dispersión 37
Momentos 47
Frecuencias 49
Cuantiles 60
Coeficiente de asimetría 69
Curtosis 72
Descripciones numéricas para datos agrupados 74
▪ Resumen de Fórmulas 77
▪ Ejercicios 78

3. Descripciones gráficas **97**
Gráfica de barras 98
Histograma 102
Polígono de frecuencias 103

Polígono de frecuencias acumuladas 104
Ojiva 105
Gráfica de pastel 106
Gráfica de tallo y hojas 110
Diagrama de caja y brazos 113
Función de distribución empírica 117
▪ Ejercicios 122

4. Descripciones para datos conjuntos **131**
Frecuencias para datos conjuntos 134
Covarianza 136
Coeficiente de correlación 140
Recta de regresión 144
Gráfica Q-Q 146
▪ Resumen de Fórmulas 151
▪ Ejercicios 151

Apéndices **156**
Notación y términos matemáticos 157
Breve orientación sobre `R` 161
Descripciones numéricas para variables aleatorias 163
Sugerencias a los ejercicios 169

Bibliografía **195**

Índice analítico **197**

Capítulo 1

Conceptos elementales

En este primer capítulo vamos a explicar los primeros conceptos elementales de la estadística descriptiva. Estudiaremos las nociones de población, muestra y variable. Explicaremos también algunas formas de clasificar las variables y mencionaremos las escalas de medición que se pueden usar para cada variable.

Población y muestra

Cotidianamente el término población se usa para referirse a un determinado grupo de personas o seres vivos. Mediante la siguiente definición ampliaremos su significado en la estadística: por población entenderemos un conjunto arbitrario de objetos. Estos objetos deberán tener ciertas características de acuerdo al estudio que deseemos llevar a cabo.

Una **población** es un conjunto de personas, objetos o eventos, de los cuales nos interesa estudiar algunas de sus características.

En un estudio estadístico la población debe especificarse los más completamente posible dependiendo de lo que se desee o se pueda estudiar u observar, y de cómo sea posible medir las características de nuestro interés. Como ejemplos de posibles poblaciones para algún estudio tenemos los siguientes: un conjunto de personas mayores a 18 años que son fumadoras, un conjunto de artículos producidos por una maquinaria, un conjunto de velocidades a

las que viajan automovilistas en un cierto punto de una avenida, un conjunto de los números de semillas en las naranjas de cierta especie cultivadas en una región en una cierta temporada, un conjunto de votantes en una elección, un conjunto de pacientes con una cierta enfermedad, etc. Como puede verse el concepto de población es realmente muy amplio.

Para un estudio estadístico, además de tener definida una población, es conveniente establecer también una unidad de observación.

Una **unidad de observación** es un grupo de elementos de una población, de la cual se tiene, o es posible obtener, su información de manera conjunta.

Así, las unidades de observación son aquellas personas, objetos, o grupos de éstos, sobre los cuales es posible obtener información para llevar a cabo el estudio estadístico. La determinación de la unidad de observación dependerá del problema a tratar y de la manera en la que la información pueda ser obtenida o que esté disponible. Por ejemplo, en un análisis cuantitativo sobre los resultados de un proceso electoral, la información puede estar disponible por casillas electorales, y en este caso las casillas electorales (grupos de votantes) pueden ser consideradas como las unidades de observación. En contraparte, si el estudio trata acerca de la intención del voto previo a la elecciones, entonces cada votante puede ser considerado como una unidad de observación.

Observemos que es posible considerar también a la población en un estudio como la totalidad de las unidades de observación, sean estas entes individuales o agrupamientos.

Nos interesa conocer ciertas características de una población y al ejercicio cuando se llevan a cabo mediciones en toda la población se le llama **censo**. En este caso el análisis estadístico y sus conclusiones se refieren a la población completa. Sin embargo, por muy diversas razones (económicas, técnicas, imposibilidad, etc.) no es posible llevar a cabo mediciones en todos los elementos de la población, de modo que debemos escoger únicamente algunos elementos y de éstos obtener sus características. Por ejemplo, si el

proceso de control de calidad de ciertos productos involucra su destrucción parcial o total, entonces no es razonable aplicar ese proceso a la totalidad de los productos. Así, a un subconjunto tomado de la población le llamaremos **muestra**, y a las mediciones que se hagan o que se tengan de una muestra les llamaremos **datos**.

Una **muestra** es cualquier subconjunto de una población. Al número de elementos de la muestra se le llama **tamaño** de la muestra.

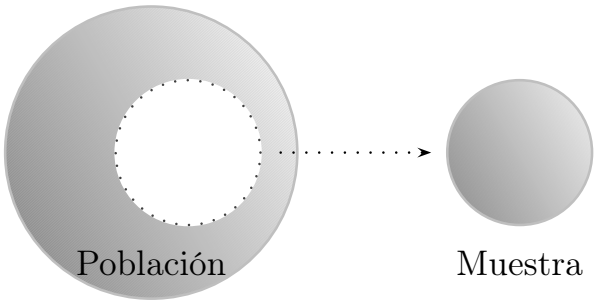


Figura 1.1: Una muestra es un subconjunto de una población.

Observemos que en el caso de un censo, la muestra está compuesta por la totalidad de la población. Además, si se concibe a la población como la totalidad de las unidades de información, entonces un muestra es cualquier colección de unidades de información.

En la Figura 1.1 se presenta de manera gráfica y simple la noción de muestra como un subconjunto de una población. Regularmente las muestras se toman mediante un mecanismo azaroso, pero tales procedimientos dependen de lo que se desee estudiar, de la forma en la que puedan medirse las variables de interés y de la estructura o agrupación que posea la población como conjunto. Reservaremos la letra n para denotar el tamaño de una muestra.

Debemos mencionar además que en ocasiones es necesario definir dos o más poblaciones para llevar a cabo estudios comparativos de ciertas caracterísi-

cas de interés, o bien obtener dos o más muestras de una misma población. También puede presentarse la necesidad de incorporar la variable tiempo en el estudio y estudiar la evolución de una característica a lo largo del tiempo.

Variables y datos

A lo que nos interesa medir y registrar en cada elemento de una población le llamaremos variable. En general, una variable es una característica que varía de un elemento a otro de la población.

Una **variable** es una característica de interés que posee cada elemento de una población y que podemos medir.

Una variable también puede considerarse como una pregunta que se le hace a cada elemento de la población, produciendo una respuesta en cada caso. Por ejemplo, en una población humana, podemos considerar la variable (pregunta) ¿Usted fuma? y obtener como respuesta "si" o "no". Para una población compuesta por un conjunto de tornillos podemos considerar la variable (pregunta) "Longitud del tornillo" y obtener como resultado de la medición un valor dentro del intervalo $(0\text{cm}, 5\text{cm})$, por ejemplo.

Mediante el término **datos** se entiende al conjunto de observaciones de una o varias variables de interés para todos los elementos de una muestra.

Generalmente un conjunto de datos se organiza y almacena en una computadora en formato de una tabla como la que se muestra en la Tabla 1.1. En esta tabla cada renglón representa una observación. En este caso tenemos a 5 personas para quienes se han registrado cuatro variables: edad, sexo, peso en kilogramos y estatura en centímetros.

Núm.	Edad	Sexo	Peso (kg.)	Estatura (cm.)
1	25	M	65	170
2	30	F	60	160
3	27	F	55	168
4	23	M	70	173
5	25	F	63	165

Tabla 1.1: Ejemplo de información tabulada.

De acuerdo al tipo de posibles respuestas que se obtengan es que las variables se pueden clasificar en varios tipos. Estudiaremos esto en la siguiente sección.

Clasificación de variables

Una primera clasificación de variables establece que éstas pueden ser cuantitativas o cualitativas. Como estos nombres lo indican, la primera se refiere a una cantidad mientras que la segunda se refiere a una cualidad.

Una variable es **cuantitativa** si sus valores son números y representan una cantidad.

Por ejemplo, el número de hijos en una familia, la longitud de un tornillo, la cantidad de desperfectos de un artículo o el número de años cumplidos son ejemplos de variables cuantitativas.

Una variable es **cualitativa** si sus valores representan una cualidad, un atributo o una categoría. Se les llama también variables **categóricas**.

Por ejemplo, la religión de una persona, su sexo, o su preferencia por algún candidato en un proceso de elección son variables cualitativas pues sus valores son atributos de las personas. El lugar de nacimiento de una persona es otro ejemplo de variable cualitativa o categórica.

Observe que se pueden usar números para etiquetar los valores de una varia-

ble cualitativa pero éstos no representan cantidades sino que se usan dichos símbolos para denotar alguna cualidad. Por ejemplo, para clasificar la calidad de un producto se pueden usar los símbolos: 2 (bueno), 1 (regular), 0 (malo). En este caso los símbolos numéricos se usan para clasificar la calidad de un producto y no se trata realmente de valores numéricos.

Regresemos a las variables cuantitativas, éstas pueden clasificarse, además, en dos categorías de acuerdo al tipo de valores que toman, pueden ser discretas o continuas. Véase la Figura 1.2.

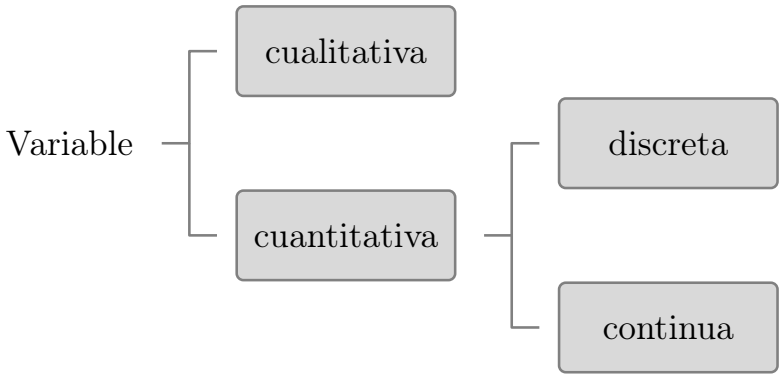


Figura 1.2: Clasificación de variables.

Una variable cuantitativa es **discreta** si el conjunto de todos sus posibles valores tiene un número finito de elementos, o bien es infinito, pero se pueden numerar uno por uno de acuerdo al conjunto de número naturales.

Por ejemplo, la colección $\{0, 1, 2, \ldots, 120\}$ puede ser el conjunto de valores de una variable cuantitativa discreta pues este conjunto tiene un número finito de elementos. Puede corresponder al número de hijos de una persona o el número de años promedio que le quedan por vivir a una persona.

Como otro ejemplo tenemos el conjunto $\{0, 1, 2, \ldots\}$, que aunque es infinito

es discreto puesto que claramente sus elementos se pueden numerar uno por uno de acuerdo al conjunto de números naturales. Los elementos de este conjunto pueden representar el número aproximado de cigarrillos que una persona fumadora ha consumido en toda su vida hasta el momento del estudio.

Una variable cuantitativa es **continua** si puede tomar todos los valores dentro de un intervalo (a, b) de números reales y no toma valores aislados.

Por ejemplo, el tiempo que le toma a una persona llegar a su lugar de trabajo o escuela puede tomar valores continuos en el intervalo $(0, \infty)$. Más generalmente, el tiempo que le toma a una persona completar una cierta actividad puede tomar este conjunto de valores.

Pueden existir variables cuantitativas cuyos valores son todos los números dentro de un intervalo (a, b) y además algunos otros puntos aislados fuera de este intervalo. Estas variables se llaman mixtas, sin embargo, por simplicidad no las consideraremos. Supondremos que nuestras variables cuantitativas son únicamente de dos tipos: discretas o continuas.

Finalmente mencionaremos que a una variable que puede tomar únicamente dos valores se le llama **variable dicotómica**. Este término se aplica tanto para variables cualitativas como cuantitativas. Por ejemplo, el sexo de una persona es una variable dicotómica pues puede tomar los valores masculino o femenino.

Escalas de medición

De acuerdo al tipo de valores que pueden tomar las variables, se pueden clasificar éstas de la siguiente manera. Para las variables cualitativas, las escalas de medición pueden ser de dos tipos: nominal u ordinal, mientras que las variables cuantitativas pueden medirse usando dos tipos de escalas: de intervalo o de razón. Explicaremos a continuación cada una de estas escalas. Empezaremos con el caso de las variables cualitativas.

Se dice que una variable cualitativa se mide mediante una **escala nominal**, o es de tipo nominal, si sus valores son etiquetas o atributos y no existe un orden entre ellos.

Por ejemplo, si nos interesa estudiar la variable cualitativa "sexo" en una población humana, sus dos posibles valores son: Masculino y Femenino. Estos dos valores son etiquetas, no existe un orden entre ellos y por lo tanto se trata de una variable de tipo nominal. Por otro lado, la variable cualitativa "Nacionalidad" también es un ejemplo de una variable de tipo nominal pues sus posibles valores: Argentina, Española, etc. son atributos y no existe un orden entre ellos. Por simplicidad consideramos en este ejemplo que cada persona tiene una única nacionalidad principal. Como un tercer ejemplo considere la variable cualitativa "religión", sus posibles valores son: Budista, Musulmana, Católica, etc. y es claro que corresponde a una variable de tipo nominal pues no hay ningún orden natural entre estos valores.

Veamos ahora la definición de la escala ordinal.

Se dice que una variable cualitativa se mide mediante una **escala ordinal**, o es de tipo ordinal, si sus valores son etiquetas o atributos pero existe un cierto orden entre ellos.

Por ejemplo, podemos considerar que la variable cualitativa "estado en el que se encuentra un artículo" tiene como posibles valores: Malo, Regular y Bueno. Es claro que estos valores son atributos de un artículo y que existe un cierto orden entre estos valores, por lo tanto, se trata de una variable de tipo ordinal.

Como un segundo ejemplo considere las siguientes calificaciones finales para un alumno en un curso: No Acreditado (NA), Suficiente (S), Bien (B) y Muy Bien (MB). Estos valores son etiquetas pero es claro que existe un orden entre estos valores, los hemos escrito en orden ascendente. Por lo tanto, esta variable, medida en el sentido indicado, es un ejemplo de una variable cualitativa de tipo ordinal.

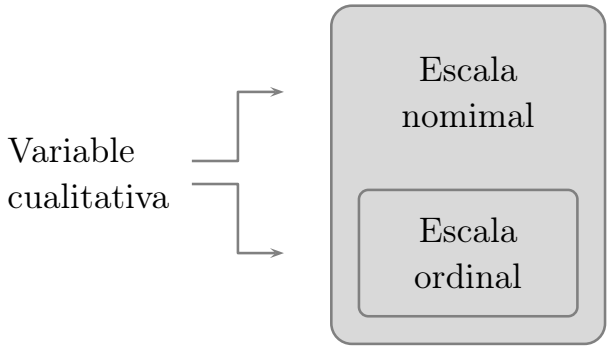


Figura 1.3: Escalas de medición para variables cualitativas.

En la Figura 1.3 se muestran gráficamente los dos tipos de escala que se usan para variables cualitativas: nominal y ordinal. Observe la contención de conjuntos que se muestra en esta figura. Esta contención significa que toda variable de tipo ordinal puede considerarse como una variable de tipo nominal, ello se logra cuando no se contempla o se ignora el orden entre los valores de la variable. La consideración contraria, sin embargo, no es posible: sin información o hipótesis adicionales, no es posible crear un orden entre los valores de una variable de tipo nominal. En la sección de ejercicios se encuentran algunos otros ejemplos de variables cualitativas con escalas de medición nominal y ordinal.

Ahora consideraremos el caso de variables cuantitativas. Recordemos que éstas pueden ser discretas o continuas, sin embargo, en las siguientes definiciones no hay ninguna distinción a este respecto, son las mismas en ambos casos. También recordemos que los valores de una variable cuantitativa son números, y por lo tanto existe ya un orden entre estos valores. Agregaremos ahora algunas condiciones adicionales a los valores numéricos de una variable cuantitativa para definir dos nuevos tipos de escalas de medición: la escala de intervalo y la escala de razón. Véase la Figura 1.4 en donde se muestra la relación general que guardan estos dos tipos de escalas. Veamos primero la definición de escala de intervalo.

Se dice que una variable cuantitativa se mide mediante una **escala de intervalo** si existe una noción de distancia entre los valores de la variable, aunque no se pueden realizar operaciones numéricas y no existe necesariamente el valor natural cero.

De esta manera no sólo la relación de orden entre los valores de una variable cuantitativa, sino que dados cualesquiera dos de sus valores podemos saber la distancia entre ellos. Por ejemplo, la escala Celsius (o Fahrenheit) para medir la temperatura es de tipo intervalo, pues existe una noción de distancia entre dos temperaturas, pero claramente no existe el valor cero natural o absoluto (el cero depende de la escala que se use, la temperatura 0°C no es la misma que 0°F). Ahora veamos la definición de escala de razón.

Se dice que una variable cuantitativa se mide mediante una **escala de razón** si los valores de la variable tienen un sentido físico y existe el cero absoluto.

Por ejemplo, la variable cuantitativa (discreta) "edad en años cumplidos de una persona" tiene como posibles valores: 0,1,. . . ,150. Por cuestiones de finitud hemos considerado una edad máxima posible de 150 años. Es claro que puede considerarse que esta variable puede medirse mediante una escala de razón pues la variable puede tomar el valor cero absoluto y existe la noción física del lapso de 1 año entre un valor y el siguiente en esta escala de medición.

Como un segundo ejemplo considere la variable cuantitativa (podemos suponer discreta) "peso" de un bebé al nacer. Puesto que siempre existe una precisión finita con la que se efectúan las mediciones, podemos considerar que el conjunto de valores de esta variable cuantitativa es un conjunto con un número finito de elementos y puede considerarse que el valor cero está incluido. Esta variable entonces se puede medir mediante una escala de razón.

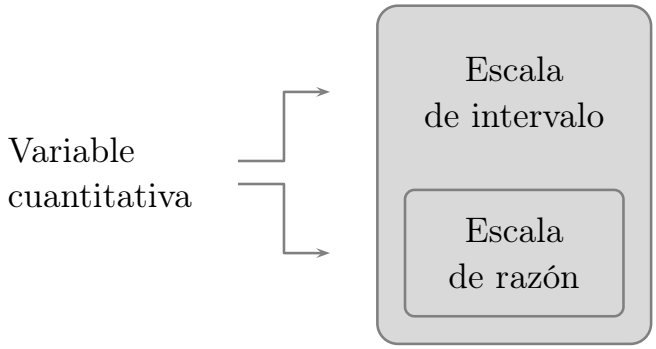


Figura 1.4: Escalas de medición para variables cuantitativas.

En la Figura 1.4 se muestran gráficamente los dos tipos de escala que se usan para variables cuantitativas. Observe nuevamente que también aquí tenemos una contención de conjuntos. Esta contención significa que toda variable con escala de medición de tipo razón puede considerarse como una variable con escala de medición de tipo intervalo, esto se consigue cuando no se contempla el sentido físico de la variable y/o no existe el cero absoluto. La consideración contraria no es posible.

Advertencia. Antes de concluir esta sección se debe mencionar que no existe una clasificación única y absoluta para una variable dada. Su tipificación dependerá del tratamiento y uso que de ella se haga. Tal vez la separación más fuerte se encuentre entre variables cualitativas y cuantitativas. De las segundas, por cuestiones de precisión numérica, una variable continua bien puede considerarse discreta.

Agrupamiento de valores

Para una variable cualitativa cualquiera tenemos una cierta cantidad de categorías $C_1, C_2, \ldots, C_k$ como sus posibles valores. Estas categorías pueden agruparse en un número menor de categorías uniendo algunas de las categorías originales. Por otro lado, para variables cuantitativas se pueden agrupar sus valores en grupos de valores $C_1, C_2, \ldots, C_k$ (estamos usando la misma notación que para el caso de categorías). Estos grupos de valores deben ser

excluyentes y exhaustivos. Esto significa que cada valor de la variable se clasifica en uno (exhaustividad) y sólo uno de estos grupos (exclusión). En cualquier caso, a los agrupamientos resultantes les llamaremos clases.

Una **clase** es un agrupamiento de categorías en el caso de variables cualitativas, o de valores numéricos en el caso de variables cuantitativas.

Observemos que las clases pueden constar de una o varias categorías en el caso de variables cualitativas, o de uno o varios números en el caso de variables cuantitativas.

Observemos además que al hacer este tipo de agrupamientos se puede perder información y el tipo de variable puede cambiar su tipo. Por ejemplo, la variable "Salario de un trabajador", que pudo ser considerada originalmente como cuantitativa, se puede transformar en una variable con valores C_1 ="Salario bajo", C_2 ="Salario medio" y C_3 ="Salario alto", la cual sería ahora una variable cualitativa de tipo ordinal.

Supongamos entonces que los valores de una variable se agrupan en clases. Al llevar a cabo una observación (u obtener un dato) de la variable, ese valor observado pertenece a una de las clases definidas y se dice entonces que la clase correspondiente fue observada.

También es posible un proceso contrario. Esto es, hay situaciones en donde los datos disponibles están agrupados en clases y no se tienen las observaciones individuales. Si por alguna razón se desea contar con los datos individuales, estos pueden obtenerse únicamente de manera aproximada eligiendo un dato representante de cada clase. A estos representantes se les llama marcas de clase.

Una **marca de clase** es un dato que representa a una clase.

Por ejemplo, si una determinada clase es un intervalo $[a, b]$ del conjunto de número reales, entonces una posible marca de clase puede ser el punto medio del intervalo, en este caso es el valor $(a + b)/2$. Véase la Figura 1.5.

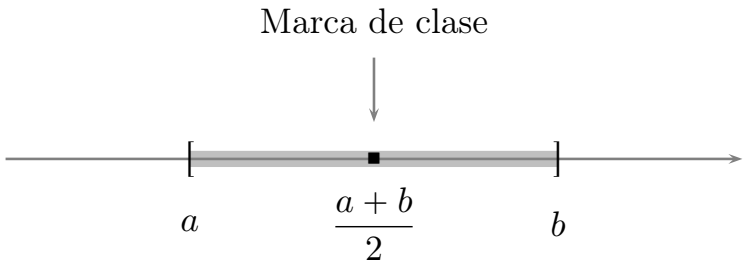


Figura 1.5: Ejemplo de marca de clase.

Cada registro o dato que se tenga de un clase se reemplaza por su representante y de esta manera se genera una colección de valores individuales aproximados de la variable. Por ejemplo, en la información que se muestra en la Tabla 1.2 aparecen clases o agrupamientos de edades en años cumplidos y una elección de las marcas de clase o dato representativo.

Clase	Marca de clase
{0,1,. . . ,9}	5
{10,11,. . . ,19}	15
⋮	⋮
{90,91,. . . ,99}	95

Tabla 1.2: Ejemplos de marcas de clase.

No existe un método estándar para determinar las marcas de clase. Cada caso debe analizarse por separado y, si es necesario, justificar la elección de las marcas de clase. Más adelante tendremos oportunidad de analizar algunos ejemplos particulares.

Sobre los datos

Muy probablemente la información que se use para llevar a cabo un estudio estadístico se encuentre almacenada en una computadora, o bien se planee capturarla en una computadora para su análisis. Aquí es útil saber que los lenguajes de programación y los programas computacionales para el análisis estadístico utilizan ciertos formatos para almacenar los valores de una variable. Cuando se define una variable en uno de estos lenguajes o programas, se le asocia, implícita o explícitamente, un tipo de dato. Como ejemplos de tipos de datos en computación tenemos los siguientes: número entero, número real, booleano, caracter, alfanumérico, etc.

Cada tipo de dato especifica los posibles valores que la variable puede tomar, su significado, las operaciones permitidas que pueden efectuarse sobre el dato y las posibles operaciones permitidas entre dos o más datos de un mismo tipo o de tipo diferente. El tipo de dato especifica además la manera en la que la información es almacenada. Otros ejemplos de tipos de datos construidos a partir de los anteriores son los siguientes: cadenas, vectores, matrices, conjuntos, etc.

Existen muchos otros tipos de datos predefinidos en los distintos programas de cómputo y existe también la posibilidad de que el usuario cree nuevos tipos de datos. Lo que queremos llamar la atención aquí es que se debe tener cuidado en el correcto manejo computacional de la información, pues dependiendo del tipo de dato declarado para cada variable es que se podrán o no se podrán hacer ciertas operaciones. Los programas de cómputo pueden efectuar operaciones sin avisar al usuario de alguna posible inconsistencia. Por ejemplo, si un dato se declara de tipo entero y se le asigna el valor 3.4, es muy posible que al efectuar alguna operación el valor se redondee a 3, sin ningún aviso al respecto. Como otro ejemplo tenemos la sucesión de caracteres “5826224”, la cual puede representar un número entero o una sucesión de siete caracteres alfanuméricos. Su tratamiento computacional dependerá del tipo de dato declarado. Sin embargo, pueden efectuarse cambios de tipo de datos para lograr algún cálculo u objetivo particular sobre el manejo del dato.

Los ejercicios que se proponen al final de cada capítulo del presente trabajo

contemplan, si acaso, muy pocos datos. Recordemos que nuestro objetivo principal es comprender los conceptos estadísticos y su posible interpretación, dejando de lado el tratamiento computacional de los datos. En situaciones reales la cantidad de información puede ser realmente enorme, y generalmente los datos están almacenados en un archivo de computadora creado con una cierta estructura de acuerdo a los datos que se ha decidido almacenar. A estos archivos se les llama **bases de datos**. Aquí es conveniente advertir que tales bases de datos pueden contener datos faltantes y algunos datos pueden ser aberrantes, es decir, tener un valor fuera de un cierto rango de valores razonable. En este trabajo consideraremos ejemplos con muy pocos datos y por lo tanto no tendremos problemas por situaciones de falta de datos o de inconsistencias en la información. En situaciones reales es muy recomendable verificar el buen estado de las bases de datos. Por ejemplo, se debe considerar la posibilidad de que haya datos faltantes y alguna decisión debe tomarse al respecto. Por problemas de captura o registro, pueden presentarse también datos erróneos o inválidos y esto afecta directamente al análisis que se realice de ellos. La insuficiente precisión o redondeo involuntario de los datos numéricos es otra variable que puede inducir errores en el análisis que se lleva a cabo y en las conclusiones que se obtengan. El tipo de dato computacional que se defina para el registro de los valores de una variable y las operaciones que el sistema computacional pueda realizar sobre ese tipo de dato es otro elemento que se debe tener presente en el análisis computacional de la información. Mencionaremos nuevamente aquí el caso de un dato con valor, por ejemplo, “435”. Esto puede representar el número indicado, o bien una cadena de tres caracteres alfanuméricos. El tratamiento que se haga de este dato puede ser diferente en cada caso.

¿Qué es la estadística descriptiva?

Con los elementos anteriores podemos ahora dar una definición aproximada de lo que trata la estadística descriptiva.

La **estadística descriptiva** es un conjunto de técnicas y procedimientos estadísticos que ayudan a describir, mostrar y resumir, la información de un conjunto de datos.

Las técnicas de la estadística descriptiva ayudan a visualizar la información de una manera más significativa, especialmente cuando la cantidad de datos es grande. Por ejemplo, puede ser más útil y más fácil de comprender la información de que en un cierto país el 52 % de la población es mujer y el 48 % es hombre. Proveer únicamente el listado completo de esta característica para cada individuo residente en este país es proporcionar demasiado detalle y eso puede ser poco útil. Como veremos más adelante, estas descripciones resumidas de la información se llevan a cabo a través de números, funciones, tablas y diagramas. Las conclusiones que pudieran obtenerse se refieren únicamente a la muestra (o muestras) en estudio y no se busca extrapolar esas conclusiones a la totalidad de la población, aunque si la muestra es "representativa" de la población, la información de la muestra puede dar indicios de la situación en la población completa.

En cambio, en la **estadística inferencial** se busca extender la información obtenida de una muestra a la población completa. En tales procedimientos no existe seguridad completa de la validez de las conclusiones y éstas pueden enunciarse únicamente con un cierto grado de confiabilidad y nunca con total seguridad. Es claro que en la estadística inferencial es crucial contar con una muestra representativa (en algún sentido) de la población. Saber obtener muestras representativas para cada situación de estudio no es una tarea fácil.

Nuestro objetivo en este trabajo es menos ambicioso y no nos ocuparemos de estudiar los métodos y procedimientos para obtener muestras, ni las maneras de extender las conclusiones obtenidas de una muestra. Supondremos simplemente que contamos con los datos de una muestra y que deseamos hacer una descripción numérica y gráfica de esa información.

Ejercicios

Población y muestra

1. Defina con la mayor precisión posible una población adecuada en el caso de que se desee llevar a cabo un estudio estadístico sobre los siguientes temas. En cada caso determine una unidad de observación.

   a) La preferencia electoral en unos comicios.

b) La calidad del aire en una ciudad.

c) Los efectos del tabaco en la salud de las personas.

d) El uso de la biblioteca en una escuela.

e) El nivel de pobreza económica de las familias en una región geográfica.

f) La calidad de los productos de una fábrica.

g) Los efectos de un nuevo medicamento para tratar una enfermedad.

h) El consumo de drogas en una ciudad.

i) La calidad del agua que llega a las casas en una ciudad.

j) El nivel de cultura de las personas en un país.

k) La relación entre pobreza y mortalidad.

l) Los niveles de audiencia de un programa de televisión.

m) La práctica de actividades deportivas en los niños.

n) La calidad del sueño en las personas en una ciudad.

ñ) La alimentación de las personas de edad avanzada.

Variables y datos

2. Defina una población adecuada, una unidad de observación y tres variables que pueden ser de interés en el caso de que se desee llevar a cabo un estudio estadístico sobre los siguientes temas.

   a) La obesidad de las personas en un país.

   b) Los niveles de inseguridad en una ciudad.

   c) El desempeño académico de estudiantes en una universidad.

   d) La situación económica de las personas de edad avanzada.

   e) La situación migratoria de extranjeros en un país.

   f) El tipo de transporte público utilizado por las personas en una ciudad.

   g) El consumo de alimentos poco saludables en los niños.

h) Las características físicas de los recién nacidos.

3. Determine una posible variable que tenga como uno de sus valores el indicado en cada inciso.

a) Descompuesto.
b) Mango.
c) Código postal 04530.
d) Obrero.
e) Promedio escolar 9.5 .
f) 5 infracciones de tránsito.
g) Ruso.
h) 25 años.
i) 3 veces.
j) 2 automóviles.
k) Rodríguez.
l) 7 de la mañana.
m) Primavera.
n) Alto.
ñ) 8 horas.
o) 1.5 horas.
p) Azul.
q) 12 días.

Clasificación de variables

4. Clasifique las siguientes variables en cualitativas o cuantitativas, en caso de ser cuantitativas diga además si son discretas o continuas.

a) El nivel de felicidad de una persona.
b) El tiempo de vida útil de un foco.
c) El número de hijos en las familias.
d) La cantidad de agua en una presa.
e) La cantidad de dinero en una cuenta bancaria.
f) El estado civil de una persona.
g) El nivel de contaminación en el aire de una ciudad.
h) La precipitación pluvial en una región.
i) La preferencia sexual de una persona.
j) El tiempo necesario para llevar a cabo un trabajo.
k) La temperatura en un cierto lugar en un cierto día.
l) El número de mascotas por familia.

m) La causa de muerte de una persona.

n) El nivel de dominio de un idioma extranjero.

ñ) El coeficiente intelectual de una persona.

o) El nivel adquisitivo de una persona.

p) Los lugares de residencia de una persona previos al actual.

q) La actividad u oficio de una persona.

r) La capacidad de ahorro de una persona.

s) El capital de una persona en el momento de su retiro laboral.

t) El resultado de un examen de matemáticas.

u) El estado de salud de una persona.

v) La estatura promedio de un equipo de basketbol.

Escalas de medición

5. Los siguientes son ejemplos de variables cualitativas. Diga si su escala de medición puede ser nominal u ordinal.

a) El color de pelo de un perro.

b) Nivel de estudios de una persona de 18 años.

c) Estado anímico de un deportista.

d) Nivel de satisfacción de un cliente que compra un producto particular.

e) Lugar de nacimiento de una persona.

f) Marca de un automóvil.

g) Preferencia del género musical de una persona.

h) Mes de nacimiento de una persona.

i) Día de la semana con mayor número de decesos.

j) Segundo idioma más hablado por las personas en un país.

k) Estación del año con mayor número de suicidios en una región.

l) Raza de gatos en una ciudad.

m) El grado militar de un soldado.

n) La ocupación de una persona.

ñ) El país destino principal de una persona que viaja al extranjero.

o) La orientación sexual de una persona.

p) La respuesta correcta o incorrecta a una pregunta en un examen.

q) El signo zodiacal de una persona.

6. Los siguientes son ejemplos de variables cuantitativas. Diga si son discretas o continuas y si su escala de medición puede ser de intervalo o de razón.

a) Número de padres vivos de una persona.

b) Talla de calzado de una persona.

c) Porcentaje de aprobación de una persona en un cargo público.

d) Peso en kilogramos de recién nacidos.

e) Número de individuos restantes de una especie en peligro de extinción.

f) La altura de los árboles de cierta especie en un bosque específico.

g) Consumo de energía eléctrica por casa en una ciudad en un periodo determinado.

h) Nivel de partículas contaminantes en una ciudad a una hora determinada.

i) Nivel de coeficiente intelectual de una persona.

j) Capital de una persona en su cuenta de ahorro para el retiro.

k) Número de días inhábiles en un calendario escolar.

l) Número de veces que una persona se cepilla los dientes al día.

m) Número de mascotas en un hogar.

n) El monto solicitado a un banco por una persona para un crédito hipotecario.

ñ) El promedio de calificaciones de un alumno.

Agrupamiento de valores

7. Suponga que los valores de la variable "Edad en años cumplidos" para una persona se agrupan en cinco categorías:

$$\begin{aligned} A &= \text{"Niño"}, \\ B &= \text{"Adolescente"}, \\ C &= \text{"Persona joven"}, \\ D &= \text{"Persona de edad media"}, \\ E &= \text{"Persona de edad avanzada"}. \end{aligned}$$

   Asigne un rango de valores para cada categoría y determine el tipo de variable así creada. Señale además su escala de medición. Si tuviera que escoger un representante (marca de clase) de cada categoría, ¿cuál sería?

8. Suponga que las letras SM denotan el salario mínimo que un trabajador gana en un cierto país. Se hace una clasificación de los salarios de todos los trabajadores de acuerdo a las siguientes tres categorías:

$$\begin{aligned} \text{Salario bajo} &= \text{"Salario desde un SM y hasta antes de 5 veces el SM"}, \\ \text{Salario medio} &= \text{"Salario desde 5 veces el SM y hasta antes de 10 veces el SM"}, \\ \text{Salario alto} &= \text{"Salario desde 10 veces el SM en adelante"}. \end{aligned}$$

   Determine el tipo de variable creada y su escala de medición. Asigne un representante (marca de clase) para cada categoría.

9. Suponga que una cierta variable numérica toma valores en el conjunto $\{1, 2, \ldots, 10\}$. Suponga que se agrupan estos valores en dos categorías: VP (valores pares) y VI (valores impares). Determine el tipo de variable creada mediante este agrupamiento y su escala de medición. Justifique la asignación de un representante (marca de clase) para cada categoría.

10. Suponga que los valores de la variable "Nacionalidad" de una persona se agrupan de acuerdo al continente al que el país pertenece. Determine el tipo de variable así creada y su escala de medición. Justifique la asignación de un representante (marca de clase) para cada categoría.

Capítulo 2

Descripciones numéricas

En este capítulo estudiaremos varias fórmulas que tienen como objetivo resumir o representar en uno, o varios números, la información de un conjunto de datos, principalmente numéricos. Supondremos que tenemos un conjunto de n mediciones

$$x_1, \ldots, x_n,$$

que representan valores observados de cierta variable de interés. Existen varias formas de resumir la información de esta colección de datos. Primeramente estudiaremos algunas cantidades que tienen como objetivo buscar un valor central que represente a los datos. Esta es la razón por la cual a estas cantidades se les llama **medidas de tendencia central** o **medidas de localización**. Estas cantidades son: la media, la moda y la mediana.

Medidas de localización

- Media
- Moda
- Mediana

Definiremos también algunas medidas de dispersión, esto es, cantidades que buscan medir algún grado de dispersión o separación entre los datos. Estudiaremos la varianza, la desviación estándar, la desviación media y el rango.

Medidas de dispersión	▪ Varianza ▪ Desviación estándar ▪ Desviación media ▪ Rango

Otras cantidades que también estudiaremos en este capítulo y que ayudan a resumir la información de un conjunto de datos son las frecuencias y los cuantiles. Véase la tabla de la página 77, en donde se resumen las cantidades que definiremos en este capítulo. Empecemos entonces con las medidas de localización.

Medidas de localización

Media

La media o media aritmética es simplemente el promedio de los números $x_1, \ldots, x_n$, esto es, se suman todos estos datos y se divide entre n. A la cantidad resultante la denotaremos por $\bar{x}$ (se lee x barra).

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

La media es la medida de localización más utilizada. Por ejemplo, supongamos que tenemos el siguiente conjunto de $n = 6$ datos que representan estaturas de personas medidas en metros:

Estaturas en metros	
$x_1 = 1.65$	$x_4 = 1.70$
$x_2 = 1.70$	$x_5 = 1.85$
$x_3 = 1.71$	$x_6 = 1.80$

La media es el promedio de estos valores:

$$\begin{aligned} \bar{x} &= \frac{1.65 + 1.70 + 1.71 + 1.70 + 1.85 + 1.80}{6} \\ &= 1.735\,. \end{aligned}$$

Este cálculo puede llevarse a cabo en el paquete estadístico `R` usando la función `mean()` como se muestra a continuación.

R

```
> x <- c(1.65,1.70,1.71,1.70,1.85,1.80)
> mean(x)
[1] 1.735
```

La media (aritmética) de un conjunto de números es entonces un valor promedio que resume y representa al conjunto de datos dado. Se le puede considerar como un representante promedio del conjunto de datos, aunque no necesariamente coincide con uno de ellos. En el ejemplo mostrado, el valor $\bar{x} = 1.735$ es la estatura promedio del conjunto de personas considerado. Observe que ninguna de estas personas tiene esta estatura.

Para un conjunto de datos pequeño como el mencionado, el cálculo de la media puede hacerse con una calculadora. Sin embargo, para un conjunto de datos más grande es conveniente tener la lista de números en un archivo de computadora, por ejemplo en una hoja de cálculo, para su cómputo expedito. En el caso de las hojas de cálculo, existen funciones similares a la mostrada para `R` para el cálculo de la media.

La media de un conjunto de datos puede interpretarse como el centro de gravedad de las observaciones cuando éstas son consideradas como pesos. Esto es, supongamos que sobre el eje horizontal se coloca un peso de cierta magnitud en cada uno de los valores observados $x_1, \ldots, x_n$. Entonces la media de estos datos es el punto de equilibrio de los pesos. Véase la Figura 2.1 en donde se muestra un ejemplo gráfico de esta interpretación de la media. Los datos son $-1, -1, -1, 0, 1, 1, 2, 2, 2, 3$ y la media es 0.8 .

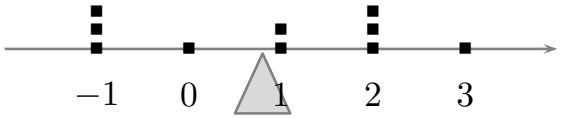


Figura 2.1: La media como punto de equilibrio de pesos.

En ocasiones las n observaciones numéricas de una variable se encuentran registradas de la siguiente forma: k valores observados distintos $x_1, \ldots, x_k$ junto con las frecuencias con las que se han registrado estos valores. Estas frecuencias $f_1, \ldots, f_k$ son números enteros mayores o iguales a cero y la suma de todas ellas es igual al tamaño de la muestra n. La media se calcula como hemos indicado antes pero en este caso se reduce a la siguiente expresión

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

Por ejemplo, supongamos que se tiene una serie de 10 observaciones de los valores -1, 0 y 1 con frecuencias 3, 5 y 2, respectivamente. La media es entonces

$$\bar{x} = \frac{1}{10} \left(-1 \cdot (3) + 0 \cdot (5) + 1 \cdot (2)\right) = -0.1\,.$$

A continuación mencionaremos algunas propiedades generales sobre el cálculo de la media bajo ciertas transformaciones de los datos.

- Supongamos que a cada uno de los datos numéricos $x_1, \ldots, x_n$ se le suma una misma cantidad c. Esta constante c puede ser positiva, negativa o cero. Definamos el nuevo conjunto de datos $y_i = x_i + c$, para $i = 1, \ldots, n$. Entonces la media del conjunto de datos modificados $y_1, \ldots, y_n$ es igual a $\bar{x}$ más la constante c. En efecto, las siguientes

operaciones aritméticas comprueban esta afirmación.

$$\begin{aligned}
\bar{y} &= \frac{1}{n}(y_1 + \cdots + y_n) \\
&= \frac{1}{n}[(x_1 + c) + \cdots + (x_n + c)] \\
&= \frac{1}{n}[(x_1 + \cdots + x_n) + (c + \cdots + c)] \\
&= \frac{1}{n}(x_1 + \cdots + x_n) + c \\
&= \bar{x} + c.
\end{aligned}$$

- Supongamos ahora que a cada uno de los datos numéricos $x_1, \ldots, x_n$ lo multiplicamos por una misma cantidad a, la cual puede ser positiva, negativa o cero. Definamos nuevamente $y_i = ax_i$, para $i = 1, \ldots, n$. Entonces la media del conjunto de datos modificados $y_1, \ldots, y_n$ es igual al producto $a\bar{x}$. En efecto,

$$\begin{aligned}
\bar{y} &= \frac{1}{n}(y_1 + \cdots + y_n) \\
&= \frac{1}{n}(ax_1 + \cdots + ax_n) \\
&= \frac{1}{n}a(x_1 + \cdots + x_n) \\
&= a\frac{1}{n}(x_1 + \cdots + x_n) \\
&= a\bar{x}.
\end{aligned}$$

Concluimos esta sección reiterando que la media es la medida de localización más utilizada, y muchas decisiones importantes son tomadas con base en esta cantidad. En el presente trabajo usaremos la media como punto de referencia al calcular algunas medidas de dispersión.

Moda

Ahora definiremos el concepto de moda. A diferencia de la media, la moda se puede calcular tanto para valores numéricos como no numéricos.

La **moda** es el valor que aparece con mayor frecuencia en el conjunto de datos, si lo hubiera.

Por ejemplo, consideremos el siguiente conjunto de edades de 6 personas.

Edades en años	
$x_1 = 15$	$x_4 = 20$
$x_2 = 20$	$x_5 = 25$
$x_3 = 15$	$x_6 = 15$

La moda es el dato 15 pues éste aparece con más frecuencia que los otros datos. Para calcular la moda de un conjunto de datos en el paquete estadístico R, no existe una función predefinida, sin embargo, pueden usarse los siguientes comandos.


```
> x <- c(15,20,15,20,25,15)
> names(sort(-table(x)))[1]
[1] ''15''
```

Consideremos ahora un ejemplo de una variable cualitativa que mide las condiciones de un producto y que tiene como posibles valores: Malo, Regular, Bueno. Suponga que tenemos el siguiente conjunto de 6 observaciones:

Condiciones de un producto	
x_1 =Malo	x_4 =Regular
x_2 =Bueno	x_5 =Malo
x_3 =Bueno	x_6 =Bueno

Entonces es claro que la moda es el valor "Bueno". En R el cálculo anterior se lleva a cabo de manera similar al caso de valores numéricos.

R

```
> x <- c(''Malo'',''Bueno'',''Bueno'',''Regular'',''Malo'',
''Bueno'')
> names(sort(-table(x)))[1]
[1] ''Bueno''
```

De esta forma, la moda es una medida de tendencia central de los datos pues indica el valor observado con mayor frecuencia. No existe una notación estándar para la moda. Se puede usar, por ejemplo, la expresión Moda(x).

Cuando se tiene un conjunto pequeño de datos es fácil determinar una posible moda, sin embargo, cuando se tiene una gran cantidad de datos, no es sencillo determinar a simple vista una posible moda. En ocasiones, ni siquiera podemos mirar los datos pues éstos se encuentran en un archivo de computadora y son tantos que no es posible pasar la mirada uno por uno.

Sobre el cálculo de la moda tenemos las siguientes observaciones:

- La moda puede no existir, es decir, puede no haber un dato con frecuencia mayor al resto de los datos. Esta situación se presenta, por ejemplo, si todos los datos son diferentes. En este caso se dice que el conjunto de datos no tiene moda.
- La moda puede existir y ser única, como en el ejemplo anterior sobre las edades. En este caso se dice que el conjunto de datos es **unimodal**.
- Pueden existir dos o más modas, es decir, pueden existir dos o mas valores que aparecen con la misma frecuencia máxima en el conjunto de datos. En este caso se dice que el conjunto de datos es **bimodal** o **multimodal**, según sea el caso.
- La moda puede permanecer sin cambio cuando se añaden u omiten datos cuya frecuencia es baja dentro del conjunto de datos.
- La moda se preserva bajo transformaciones lineales. Esto eso, si Moda(x) es una moda del conjunto de datos numéricos $x_1, \ldots, x_n$, y si se define $y_i = ax_i + c$ para $i = 1, \ldots, n$, con a y c dos constantes cualesquiera, entonces una moda de $y_1, \ldots, y_n$ es

$$\text{Moda}(y) = a \cdot \text{Moda}(x) + c.$$

Como hemos mostrado, la moda puede calcularse para cualquier tipo de datos, sean éstos cualitativos o cuantitativos. Además, en el caso de tener datos agrupados, se puede calcular la moda de estas clases o categorías y se pueden usar los términos “”clase modal” o “intervalo modal”, según sea el caso. En la Figura 2.2 se muestran dos ejemplos gráficos de la moda.

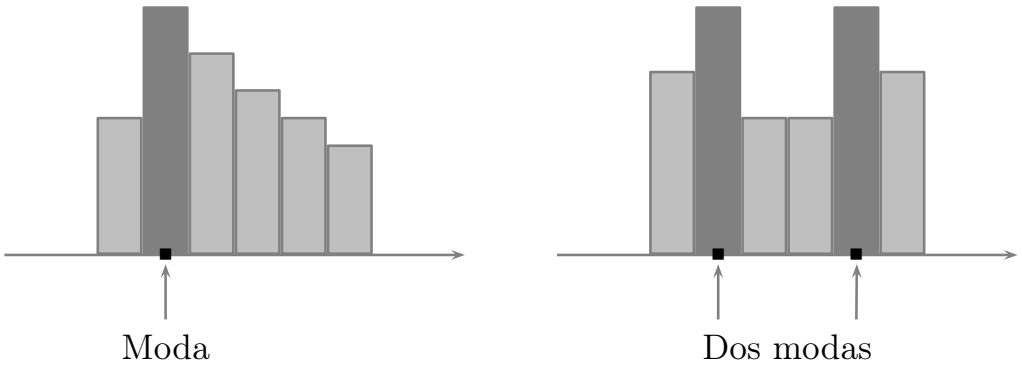


Figura 2.2: Ejemplos gráficos de modas.

Mediana

Otra medida de tendencia central para datos numéricos es la mediana. Supongamos nuevamente que tenemos una colección de n números

$$x_1, \ldots, x_n.$$

Podemos ordenar estos números de menor a mayor, incluyendo repeticiones, y obtener la colección

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)},$$

en donde $x_{(1)}$ denota el número más pequeño, $x_{(2)}$ denota el segundo número más pequeño, etcétera, hasta $x_{(n)}$ que denota el número más grande. Es claro que algunos de estos números pueden repetirse, es decir, algunos de estos datos pueden aparecer varias veces en esta ordenación. En este procedimiento es importante conservar estas repeticiones. La mediana se calcula de la siguiente forma.

La **mediana** es el dato ordenado de en medio, esto es,

- Si el número de datos n es par, entonces existen dos datos ordenados de en medio y la mediana es el promedio de estos dos números, esto es $(x_{(n/2)} + x_{(n/2)+1})/2$.
- Si el número de datos n es impar, entonces el dato ordenado de en medio es $x_{(n-1)/2}$ y esta es la mediana.

Figura 2.3: Ejemplos gráficos del cálculo de la mediana.

En la Figura 2.3 se ilustra gráficamente el cálculo de la mediana usando unos pocos datos representados por puntos distintos sobre el eje real. Se muestra el caso cuando el número de datos es par y después cuando el número de datos es impar. De esta manera, la mediana es un valor central que separa al conjunto de datos ordenados en dos partes iguales y representa un valor central típico del conjunto de datos observado (aunque puede no ser ninguno de los valores observados). No existe una notación estándar para la mediana, en este trabajo la denotaremos por $\tilde{x}$ (se lee x tilde).

Veamos algunos ejemplos del cálculo de la mediana considerando un número pequeño de datos. Supongamos que tenemos el registro de las siguientes estaturas (en centímetros) de 6 personas:

Estaturas en centímetros					
165	172	170	165	174	182

Ordenando estos números de menor a mayor, incluyendo repeticiones, se obtiene el siguiente arreglo:

165	165	170	172	174	182

Como se trata de un número par de datos, la mediana es el promedio de los dos datos centrales, esto es,

$$\begin{aligned} \tilde{x} &= \frac{170+172}{2} \\ &= 171. \end{aligned}$$

En este caso la mediana es un valor no observado. Usando el paquete `R` la mediana se calcula mediante la función `median()` como se muestra a continuación:

R

```
> x <- c(165,172,170,165,174,182)
> median(x)
[1] 171
```

Vamos a agregar el dato 175 en el ejemplo anterior para así tener 7 datos. Tenemos ahora un número impar de datos. Los datos ordenados son

165	165	170	172	174	175	182

Como se trata de un número impar de datos, la mediana es el dato central, esto es, el dato 172. Se puede comprobar este cálculo de la mediana en el paquete `R` de la siguiente forma:

R

```
> x <- c(165,172,170,165,174,182,175)
> median(x)
[1] 172
```

De esta manera, la mediana es un número que separa a los datos (ordenados de menor a mayor) en dos partes con igual número de datos: la primera parte son los números que son menores o iguales a la mediana, y la segunda parte corresponde al conjunto de números que son mayores o iguales a la mediana.

En la Figura 2.4 se muestran dos ejemplos gráficos de la mediana.

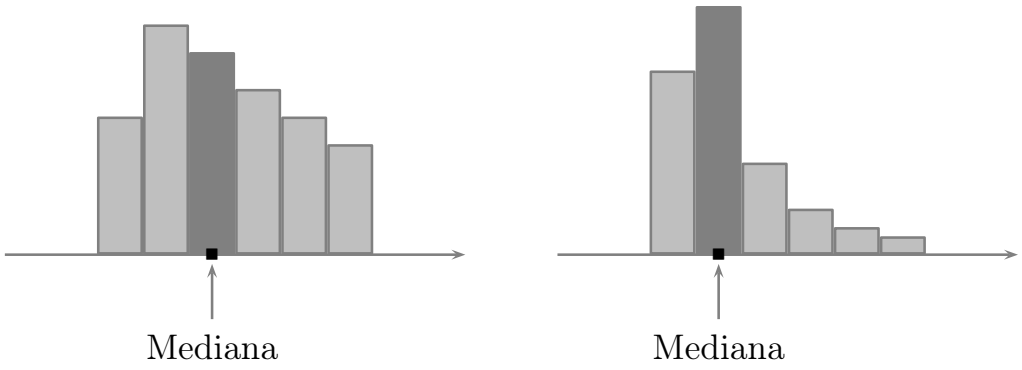


Figura 2.4: Ejemplos gráficos del cálculo de la mediana.

A partir de los ejemplos mostrados deben ser claras las siguientes observaciones acerca de la mediana. Usaremos pocos datos para ilustrar los resultados.

- La mediana puede ser uno de los datos observados o no serlo. Lo es cuando el número de datos es impar, por ejemplo,

1	1	2	2	3	4	4

  o cuando el número de datos es par y los dos datos de en medio son iguales, por ejemplo,

1	1	2	2	2	3

La mediana no es uno de los datos observados cuando el número de datos es par y los dos datos de en medio son distintos. En el siguiente ejemplo la mediana es el dato no observado $\tilde{x} = (2+3)/2 = 2.5$.

1	2	2	3	4	5

- La mediana es insensible a cambios de algunos de los datos, siempre y cuando estos cambios se efectúen dentro de la misma mitad donde se encuentran los datos modificados. Por ejemplo, todos los siguientes conjuntos de datos difieren en posiciones dentro de la misma mitad, izquierda o derecha, de los datos ordenados. La mediana $\tilde{x} = 2$ es la misma.

1	1	2	2	3	4	4

0	1	2	2	3	4	5

1	2	2	2	4	4	4

- La mediana se preserva bajo transformaciones lineales. Esto eso, si $\tilde{x}$ es la mediana del conjunto de datos $x_1, \ldots, x_n$, y si se define $y_i = ax_i + c$ para $i = 1, \ldots, n$, con a y c dos constantes cualesquiera, entonces la mediana de $y_1, \ldots, y_n$ es

$$\tilde{y} = a\tilde{x} + c.$$

Los siguientes cálculos muestran la veracidad de esta afirmación. Supongamos que $a \geqslant 0$. Entonces los datos originales ordenados de menor a mayor $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$ se transforman en los datos ordenados

$$ax_{(1)} + c \leqslant \cdots \leqslant ax_{(n)} + c$$

Si el número n de datos es impar, entonces el dato de en medio de esta nueva colección es

$$ax_{(n-1)/2} + c = a\tilde{x} + c.$$

Si el número n de datos es par, entonces el dato de en medio es

$$\begin{aligned} \frac{(ax_{(n/2)} + c) + (ax_{(n/2)+1} + c)}{2} &= a\,\frac{x_{(n/2)} + x_{(n/2)+1}}{2} + c \\ &= a\tilde{x} + c. \end{aligned}$$

El mismo resultado se obtiene cuando $a < 0$. Así, hemos comprobado que la mediana se preserva bajo cualquier transformación lineal aplicada a un conjunto de datos.

¿Qué medida de localización es mejor?

No existe tal cosa. Cada una de las medidas de localización que hemos mencionado mide de manera diferente la centralidad de un conjunto de datos numéricos. Sin embargo, la media es la medida de localización que con mayor frecuencia se utiliza en los estudios estadísticos. Por otro lado, aunque el cálculo de la mediana puede ser más complicado, ésta tiene menor afectación ante la presencia de errores en los datos o a valores extremos de éstos.

A menudo se usan términos como "ingreso medio" o "tiempo medio de vida" sin especificar (a veces por conveniencia) la medida de localización utilizada. Ésta puede ser la media, la mediana o la moda. Para evitar confusiones y para fines de comparación, en cualquier estudio debe especificarse plenamente la forma de calcular el "valor medio" al que se refieren estas expresiones.

En gráficas de frecuencias que presentan simetría perfecta como la que se ilustra en la Figura 2.5, la media, la moda y la mediana, coinciden. En general, esta situación no siempre se presenta.

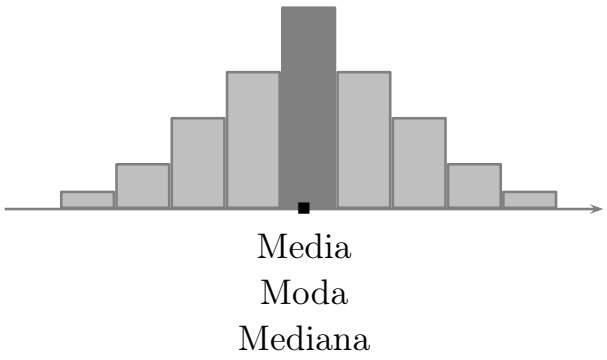


Figura 2.5: Situación donde la media, la moda y la mediana coinciden.

En la Figura 2.6 se muestran dos ejemplos en donde las tres medidas de localización que hemos definido pueden tomar valores distintos. Para la gráfica de la izquierda, los datos son: $0, 1, 2, 2$. La media es $\bar{x} = 1.25$, la mediana es $\tilde{x} = 1.5$ y la moda es 2. Se verifica que $\bar{x} < \tilde{x} < \text{Moda}(x)$.

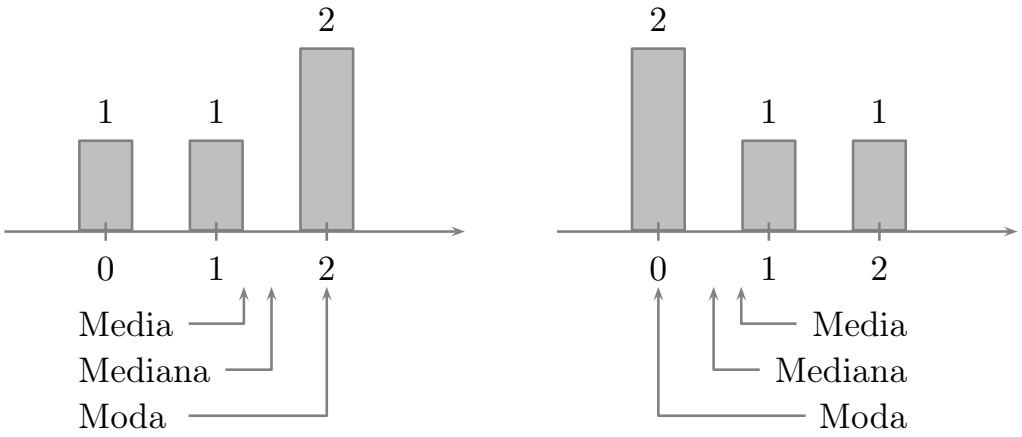


Figura 2.6: Ejemplos de media, moda y mediana.

Para la gráfica de la derecha, tenemos cuatro datos: $0, 0, 1, 2$. La media es $\bar{x} = 0.75$, la mediana es $\tilde{x} = 0.5$ y la moda es 0. En este caso tenemos el orden $\text{Moda}(x) < \tilde{x} < \bar{x}$.

Medidas de dispersión

Ahora estudiaremos algunas cantidades que permiten medir el grado de dispersión de un conjunto de datos numéricos. En casi todas estas medidas de dispersión es necesario considerar un valor central de los datos como punto de referencia. Como tal valor central tomaremos a la media $\bar{x}$. Cualquier otra medida de localización puede usarse como valor central pero, siguiendo lo mayormente usado, la media es nuestra elección.

Varianza

La varianza es un promedio de la distancia al cuadrado de cada uno de los datos x_i respecto de la media $\bar{x}$ y es la medida de dispersión más comúnmente usada. Se calcula de la forma siguiente.

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

Para especificar que se trata de la varianza de un conjunto de datos denotado por x, a la varianza la denotaremos también por los símbolos s_x^2, $s^2(x)$ o también por $\text{var}(x)$. Es claro que para calcular la varianza primero es necesario encontrar la media $\bar{x}$. Por ejemplo, consideremos el siguiente conjunto de pesos en kilogramos de 6 personas:

Pesos en kilogramos	
$x_1 = 70$	$x_4 = 66$
$x_2 = 68$	$x_5 = 70$
$x_3 = 75$	$x_6 = 65$

Puede comprobarse que la media es $\bar{x} = 69$ y por lo tanto la varianza se obtiene como sigue:

$$\begin{aligned} s^2 &= \frac{1}{6}(70-69)^2 + \frac{1}{6}(68-69)^2 + \frac{1}{6}(75-69)^2 \\ &\quad + \frac{1}{6}(66-69)^2 + \frac{1}{6}(70-69)^2 + \frac{1}{6}(65-69)^2 \\ &= 10.666\,. \end{aligned}$$

El cálculo de la varianza puede llevarse a cabo en el paquete estadístico R usando la función `var()`, como se muestra enseguida.

```
> x <- c(70,68,75,66,70,65)
> var(x)
[1] 12.8
```

Se puede observar que hay una diferencia entre el valor presentado antes y el que reporta el paquete R en el cálculo de esta varianza. La razón se encuentra en que la varianza también puede definirse como se indica en la siguiente fórmula y es la que usa R en su cálculo.

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

Puede notarse que en esta expresión aparece el denominador $n-1$ en lugar de n. Esta fórmula es usada con mucha frecuencia debido a que, cuando se aplica al caso de variables aleatorias, satisface una propiedad estadística importante llamada insesgamiento. Así, debe tenerse en cuenta esta diferencia en el cálculo de la varianza. En el presente texto usaremos la fórmula con denominador n. Esta fórmula es más natural y es consistente con otras cantidades que definiremos más adelante llamadas momentos. Para el cálculo particular mostrado en donde se tienen $n = 6$ datos, puede comprobarse que

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 &= \frac{1}{5} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \frac{6}{5} \left(\frac{1}{6} \sum_{i=1}^{n} (x_i - \bar{x})^2\right) \\ &= \frac{6}{5} \left(10.666\cdots\right) \\ &= 12.8\,. \end{aligned}$$

A continuación veremos algunas propiedades de la varianza bajo transformaciones del conjunto de datos.

- Sea s_x^2 la varianza del conjunto de datos numéricos $x_1, \ldots, x_n$. Supongamos que a cada una de estos números se le suma una misma cantidad c. Esta constante c puede ser positiva, negativa o cero. Definamos el nuevo conjunto de datos $y_i = x_i + c$, para $i = 1, \ldots, n$, y denotemos por s_y^2 a la varianza de estos números. Entonces la varianza del conjunto de datos modificados $y_1, \ldots, y_n$ es idéntica a la varianza de $x_1, \ldots, x_n$, es decir, $s_y^2 = s_x^2$. En efecto, recordando que $\bar{y} = \bar{x} + c$, tenemos que

$$\begin{aligned} s_y^2 &= \frac{1}{n}\left((y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\right) \\ &= \frac{1}{n}\left((x_1 + c - \bar{x} - c)^2 + \cdots + (x_n + c - \bar{x} - c)^2\right) \\ &= \frac{1}{n}\left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right) \\ &= s_x^2. \end{aligned}$$

  De esta manera, añadir una constante a un conjunto de datos numéricos no modifica su varianza. Esto debe ser intuitivamente claro pues la dispersión de los datos no se modifica al llevar a cabo una traslación de esa forma.

- Supongamos ahora que a cada uno de los datos numéricos $x_1, \ldots, x_n$ lo multiplicamos por una misma cantidad a, la cual puede ser positiva, negativa o cero. Definamos nuevamente $y_i = ax_i$, para $i = 1, \ldots, n$. Entonces la media del conjunto de datos modificados $y_1, \ldots, y_n$ es igual al producto $a^2 s_x^2$. En efecto, recordando que $\bar{y} = a\bar{x}$, tenemos que

$$\begin{aligned} s_y^2 &= \frac{1}{n}\left((y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\right) \\ &= \frac{1}{n}\left((ax_1 - a\bar{x})^2 + \cdots + (ax_n - a\bar{x})^2\right) \\ &= \frac{1}{n}\left(a^2(x_1 - \bar{x})^2 + \cdots + a^2(x_n - \bar{x})^2\right) \\ &= a^2 \cdot \frac{1}{n}\left((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right) \\ &= a^2 \cdot s_x^2. \end{aligned}$$

  Así, hemos comprobado que la varianza de un conjunto de datos multiplicados por una constante es igual a la constante al cuadrado multiplicada por la varianza de los datos originales.

Los dos resultados anteriores pueden combinarse para afirmar que la varianza del conjunto de datos transformados $y_i = ax_i + c$ es $s_y^2 = a^2 \cdot s_x^2$. Se pide verificar este resultado en uno de los ejercicios.

El cálculo de la varianza para datos agrupados puede efectuarse de la siguiente forma: si se tienen n observaciones de k valores distintos $x_1, \ldots, x_k$ con frecuencias $f_1, \ldots, f_k$, la varianza se reduce a la fórmula:

$$s^2 = \frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x})^2 \, f_i.$$

Desviación estándar

A la raíz cuadrada positiva de la varianza se le llama desviación estándar o desviación típica, y se le denota por la letra s. Así, para su cálculo se usa la siguiente fórmula:

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

Por ejemplo, para el conjunto de datos de pesos dados en kilogramos de 6 personas mostrados líneas arriba para el cálculo de la varianza, tenemos que la desviación estándar es

$$s = \sqrt{10.666 \cdots} = 3.265986\,.$$

Lo cual en `R` se puede calcular mediante la función `sd()`. Las letras sd provienen del término en inglés *standard deviation*.

R

```
> x <- c(70,68,75,66,70,65)
> sd(x)
[1] 3.577709
```

Debido a que `R` calcula la varianza con denominador $n - 1$, se obtiene el resultado mostrado en el recuadro. En efecto, usando `R` obtuvimos antes que

la varianza es 12.8, de modo que $\sqrt{12.8} = 3.57$.

A diferencia de la varianza, la desviación estándar posee la buena cualidad de estar expresada en las mismas unidades de medición que la de los datos observados. Para el ejemplo mostrado se trata de kilogramos, mientras que la varianza tiene unidades de medición kilogramos al cuadrado.

A continuación vamos a mencionar el cambio que tiene la desviación estándar cuando los datos observados se modifican mediante una transformación lineal. La fórmula que encontraremos es consecuencia de los resultados antes vistos para la varianza.

- Si el conjunto de datos $x_1, \ldots, x_n$ tiene desviación estándar s_x y si se define la transformación $y_i = a\,x_i + c$ para cada $i = 1, \ldots, n$, en donde a y c son dos constantes arbitrarias, entonces la desviación estándar de los datos $y_1, \ldots, y_n$ es

$$s_y = \sqrt{s_y^2} = \sqrt{a^2 \cdot s_x^2} = |a| \cdot s_x.$$

  Es decir, bajo la transformación lineal $x \mapsto y = ax + c$, la desviación estándar s_y es igual a la desviación estándar original s_x multiplicada por el valor absoluto de la constante a.

Desviación media

Al promedio de los valores absolutos de las diferencias entre los datos y la media se le llama desviación media. Más específicamente, supongamos que $\bar{x}$ es la media de los datos numéricos $x_1, \ldots, x_n$, entonces la desviación media se denota por $\mathrm{dm}(x)$ y se define como el siguiente promedio.

$$\mathrm{dm}(x) = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Así, la desviación media es otra medida de la dispersión de un conjunto de datos numéricos. Por ejemplo, consideremos nuevamente los datos del peso de seis personas.

Peso en kilogramos de seis personas	
$x_1 = 70$	$x_4 = 66$
$x_2 = 68$	$x_5 = 70$
$x_3 = 75$	$x_6 = 65$

Habíamos encontrado antes que la media para este conjunto de datos es $\bar{x} = 69$. Entonces la desviación media es

$$\begin{aligned} \text{dm}(x) &= \frac{1}{6}|70-69| + \frac{1}{6}|68-69| + \frac{1}{6}|75-69| \\ &\quad + \frac{1}{6}|66-69| + \frac{1}{6}|70-69| + \frac{1}{6}|65-69| \\ &= \frac{1}{6}(1+1+6+3+1+4) \\ &= 16/6 \\ &= 2.66\,. \end{aligned}$$

Este mismo cálculo se puede llevar a cabo en el paquete `R` mediante el siguiente código.

R

```
> x <- c(70,68,75,66,70,65)
> sum(abs(x-mean(x)))/length(x)
[1] 2.666667
```

Advertencia: existe el término desviación media absoluta (*mean absolute deviation*) que se calcula como antes pero tomando a la mediana de los datos como punto central y no la media $\bar{x}$ como lo hemos hecho aquí. En el código `R` que hemos mostrado aparece de manera explícita la fórmula utilizada en el cálculo. Si se hace uso de una computadora y se emplea alguna función predefinida para calcular la desviación media, es recomendable verificar el punto central empleado en el cálculo de la función.

Mencionaremos ahora algunas propiedades generales de la desviación media.

- Si los datos $x_1, \ldots, x_n$ se trasladan c unidades y por lo tanto se transforman en $x_1 + c, \ldots, x_n + c$, en donde c es una constante cualquiera, entonces la desviación media no cambia pues, recordando que los datos transformados tienen media $\bar{x} + c$, tenemos que

$$\begin{aligned} \mathrm{dm}(x + c) &= \frac{1}{n} \sum_{i=1}^{n} |(x_i + c) - (\bar{x} + c)| \\ &= \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \\ &= \mathrm{dm}(x). \end{aligned}$$

- Si los datos $x_1, \ldots, x_n$ se multiplican por una constante a y por lo tanto se transforman en $ax_1, \ldots, ax_n$, entonces estos nuevo datos tienen desviación media $\mathrm{dm}(ax) = |a| \, \mathrm{dm}(x)$. En efecto, recordando que la media de los datos transformados es $a\bar{x}$, tenemos que

$$\begin{aligned} \mathrm{dm}(ax) &= \frac{1}{n} \sum_{i=1}^{n} |ax_i - a\bar{x}| \\ &= \frac{1}{n} \sum_{i=1}^{n} |a| \, |x_i - \bar{x}| \\ &= |a| \cdot \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \\ &= |a| \cdot \mathrm{dm}(x). \end{aligned}$$

Rango

Ahora definiremos el rango de una colección de números $x_1, \ldots, x_n$. Para calcular esta cantidad es necesario identificar el dato más pequeño $x_{(1)}$ y el dato más grande $x_{(n)}$. El rango de la colección de números dada se denota por la letra r y es simplemente el dato mayor menos el dato menor.

$$r = x_{(n)} - x_{(1)}$$

Es claro que el rango de un conjunto de datos numéricos es una medida de dispersión, pues indica la distancia máxima entre cualesquiera dos datos. El

rango también puede interpretarse como la longitud del intervalo más pequeño en el que se encuentran todos los datos observados.

El rango de un conjunto de datos numéricos puede calcularse en R mediante el comando `range()`. La aplicación de esta función en R produce el cálculo del valor mínimo y el valor máximo del conjunto de datos numéricos dado. La diferencia de estas dos cantidades produce el número que hemos definido como rango. Aquí tenemos un ejemplo.

```
> x <- c(13,25,12,10,24,14)
> range(x)
[1] 10 25
```

Alternativamente, pueden usarse las funciones `min()` y `max()`, cuyo significado es evidente, para calcular el rango en R.

```
> x <- c(13,25,12,10,24,14)
> min(x)
[1] 10
> max(x)
[1] 25
```

Usaremos la expresión r_x, $r(x)$ o Rango(x) para denotar al rango de un conjunto de números $x_1, \ldots, x_n$, aunque esta notación no es estándar. Los siguientes resultados se pueden comprobar sin mucha dificultad y establecen algunas propiedades generales del rango.

- Sea r_x el rango del conjunto de datos $x_1, \ldots, x_n$. Defina $y_i = x_i + c$, para $i = 1, \ldots, n$, con c una constante arbitraria. Sea r_y el rango de $y_1, \ldots, y_n$. Entonces las cantidades r_y y r_x coinciden. En efecto,

$$r_y = (x_{(n)} + c) - (x_{(1)} + c) = x_{(n)} - x_{(1)} = r_x.$$

  Es decir, trasladar los datos c unidades no modifica el rango.

- Sea r_x el rango del conjunto de datos $x_1, \ldots, x_n$. Defina $y_i = a\, x_i$, para $i = 1, \ldots, n$, con $a \geqslant 0$ una constante. Sea r_y el rango de $y_1, \ldots, y_n$. Entonces, como a es mayor o igual a cero, el valor máximo del nuevo conjunto de datos es $y_{(n)} = ax_{(n)}$ y el valor mínimo es $y_{(1)} = ax_{(1)}$. Por lo tanto,

$$r_y = a\, x_{(n)} - a\, x_{(1)} = a\,(x_{(n)} - x_{(1)}) = a\, r_x.$$

  Esto significa que si los datos se multiplican por una constante a mayor o igual a cero, entonces el rango también se multiplica por esa constante. ¿Cómo cambia este resultado cuando $a < 0$?

- El rango no cambia cuando se añaden u omiten datos, siempre y cuando no se modifique el valor máximo ni el valor mínimo del conjunto de datos original.

Coeficiente de variación

Sea $x_1, \ldots, x_n$ una colección de n observaciones de una variable cuantitativa. Sea $\bar{x}$ su media y sea s su desviación estándar. Al cociente $s/\bar{x}$ se le llama coeficiente de variación y se le denota por $\mathrm{cv}(x)$, suponiendo por supuesto que $\bar{x} \neq 0$.

$$\mathrm{cv}(x) = \frac{s}{\bar{x}}$$

Tanto la desviación estándar s como la media $\bar{x}$ poseen las mismas unidades de medición. Por lo tanto, el cociente de estas cantidades no posee unidad de medición y, en consecuencia, el coeficiente de variación puede servir para comparar la dispersión de dos o más conjuntos de datos de variables cuantitativas.

Aquí tenemos dos propiedades generales del coeficiente de variación:

- Sea y el conjunto de datos transformados $y_i = a\, x_i$, en donde a es una constante distinta de cero. Esto corresponde a un cambio de escala en la medición de los datos. Hemos comprobado antes que $s_y = |a| \cdot s_x$ y

que $\bar{y} = a \cdot \bar{x}$. Por lo tanto,

$$\text{cv}(y) = \frac{|a|}{a} \cdot \frac{s_x}{\bar{x}} = \begin{cases} \text{cv}(x) & \text{si } a > 0, \\ -\text{cv}(x) & \text{si } a < 0. \end{cases}$$

Es decir, bajo cambios de escala, el coeficiente de variación puede cambiar de signo, pero no cambia de magnitud.

- Sea y el conjunto de datos transformados $y_i = x_i + c$, en donde c es una constante arbitraria. Esto corresponde a una traslación de los datos. Hemos comprobado antes que $s_y = s_x$ y que $\bar{y} = \bar{x} + c$. Por lo tanto,

$$\text{cv}(y) = \frac{s_x}{\bar{x} + c}.$$

¿Qué medida de dispersión es mejor?

No existe tal cosa. Cada una de las medidas de dispersión que hemos mencionado mide de manera diferente la variabilidad de un conjunto de datos numéricos. Sin embargo, la varianza es la que más comúnmente se utiliza en los estudios estadísticos. Recordemos además que algunas de estas medidas de dispersión se calculan respecto de una medida de localización central de los datos. En nuestras definiciones, se ha usado la media. Si alguna otra medida de centralidad como la mediana se usa en las definiciones, se obtienen otras medidas de variabilidad ligeramente distintas.

Por ejemplo, en la Figura 2.7 se muestran las gráficas de frecuencias de dos conjuntos de datos. Los datos x de la gráfica de la izquierda muestran poca variabilidad, comparados con los datos y de la gráfica de la derecha. Puede verificarse visualmente que $\bar{x} = \bar{y} = 2$, y que cada una de las medidas de variabilidad que hemos definido es menor para x que para y. Esto se muestra a continuación. Los cálculos fueron hechos en el paquete `R` y se han escrito con únicamente dos dígitos decimales.

$$\begin{aligned} s^2(x) = 0.50 &< 2.50 = s^2(y), \\ s(x) = 0.70 &< 1.58 = s(y), \\ \mathrm{dm}(x) = 0.40 &< 1.20 = \mathrm{dm}(y), \\ r(x) = 2 &< 4 = r(y), \\ \mathrm{cv}(x) = 0.35 &< 0.79 = \mathrm{cv}(y). \end{aligned}$$

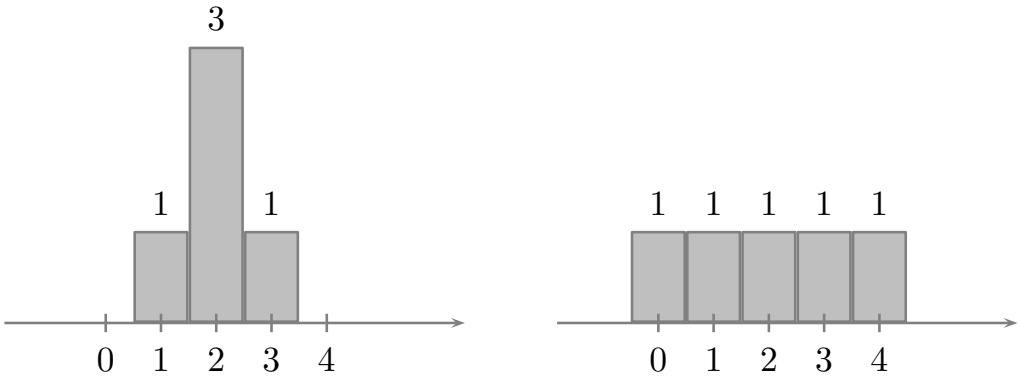


Figura 2.7: Ejemplos de dispersión grande y pequeña.

Momentos

Las cantidades que hemos definido como media y varianza pueden generalizarse a un concepto más amplio llamado momento. Consideremos una vez más que tenemos una serie de observaciones $x_1, \ldots, x_n$ de una variable cuantitativa de interés. Sea $k \geqslant 1$ un número entero. A la cantidad definida a continuación se le llama el k-ésimo momento muestral, o bien momento muestral de orden k.

$$m'_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$$

Se trata simplemente del promedio aritmético de cada uno de los datos elevados a la potencia k. El valor entero de k determina el numeral del momento,

así por ejemplo, tenemos que el primer momento ($k = 1$) es la media, el segundo momento ($k = 2$) es el promedio de los datos elevados al cuadrado, etcétera. Si x denota el conjunto de datos $x_1, \ldots, x_n$, entonces se puede usar el término $m'_k(x)$ para denotar el k-ésimo momento de x.

Veamos un ejemplo. Supongamos que x denota el conjunto de los siguientes cuatro datos: $2, 4, 1, 3$. Entonces el segundo momento es

$$m'_2(x) = \frac{1}{4}\left(2^2 + 4^2 + 1^2 + 3^2\right) = \frac{30}{4} = 7.5\,.$$

Cada momento es una medida de cierta característica de los datos. Sin embargo, no se conoce la característica que se está midiendo en cada caso, únicamente se conoce para los primeros pocos momentos. Por ejemplo, el primer momento es la media y esta es una medida de localización o centralidad de los datos, el segundo momento está relacionado con la varianza y esta es una medida de la dispersión de los datos, el tercer momento está relacionado con la asimetría de los datos, el cuarto momento está relacionado con la forma de las colas de la gráfica de frecuencias de los datos, es decir, de la manera en la que decae o se desvanece a cero la gráfica de frecuencias en sus dos extremos: izquierdo y derecho. Y esto es todo, en general no existen interpretaciones bien establecidas para los momentos de orden superior.

En el paquete `R` pueden calcularse los momentos de un conjunto de datos numéricos usando el comando `moment()`. Para ello se necesita instalar previamente el paquete `moments`. Reproduciremos el ejemplo mostrado antes, pero ahora en `R`.


```
> x <- c(2,4,1,3)
> moment(x,order=2)
[1] 7.5
```

Existen además otros tipos de momentos. Por ejemplo, si $\bar{x}$ es la media de los datos, entonces a las cantidades que se definen a continuación se les conoce como momentos centrales de orden k.

$$m_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k$$

Es decir, tenemos nuevamente un promedio aritmético pero esta vez se trata de los datos centralizados al restarles a cada uno de ellos la media $\bar{x}$. No es difícil verificar que $m_1 = 0$ y que m_2 es la varianza de los datos. Añadiendo la opción `central=TRUE` a la función `moment()` en el código `R` arriba indicado, se pueden calcular estos momentos centrales. Como hemos indicado, se pueden definir otros tipos de momentos para un conjunto de datos numéricos, pero no los mencionaremos pues no nos serán de utilidad en este trabajo.

Frecuencias

En las siguientes secciones estudiaremos varios tipos de frecuencias que se pueden definir para un conjunto de datos. Algunas de estas frecuencias pueden definirse tanto para variables cualitativas como para variables cuantitativas.

Frecuencias absolutas

Supongamos que $C_1, C_2, \ldots, C_k$ representan las categorías de una variable cualitativa, o bien agrupamientos excluyentes y exhaustivos de los valores de una variable cuantitativa. Recordemos que a estas categorías o agrupamientos les hemos llamado clases, y la letra C ayuda a recordar su significado. Para variables cuantitativas, estas clases pueden contener varios valores o solamente un valor. En el segundo caso, estamos en la situación en donde se está considerando un número finito de valores $x_1, \ldots, x_k$ de una variable cuantitativa discreta.

Supongamos además que al hacer n observaciones de la variable, contamos con el número de veces que cada una de estas clases fue observada:

$$\begin{aligned} &\text{La clase } C_1 \text{ fue observada } f_1 \text{ veces,} \\ &\text{La clase } C_2 \text{ fue observada } f_2 \text{ veces,} \\ &\qquad\vdots \qquad\qquad\qquad\vdots \\ &\text{La clase } C_k \text{ fue observada } f_k \text{ veces.} \end{aligned}$$

Las cantidades $f_1, \ldots, f_k$ son entonces números enteros mayores o iguales a cero que indican las frecuencias con las que fueron observadas cada una de las clases. A estas cantidades se les llama frecuencias absolutas o simplemente frecuencias, y la letra f proviene de la primera letra de este término. Como se tienen n observaciones de la variable, tenemos que

$$f_1 + \cdots + f_k = n.$$

Esta información puede representarse en forma tabular como se muestra en la Tabla 2.1. En esta tabla están representadas todas las clases consideradas y sus respectivas frecuencias. Esta es una manera muy útil de resumir los datos. En el siguiente capítulo haremos gráficas de estas frecuencias y ello ayudará visualmente a tener una mejor comprensión de la distribución de los datos.

Clase	Frecuencia (absoluta)
C_1	f_1
C_2	f_2
$\vdots$	$\vdots$
C_k	f_k
Suma	$n = f_1 + \cdots + f_n$

Tabla 2.1: Ejemplo general de clases y sus frecuencias.

La **frecuencia** de una clase (categoría o conjunto de valores) es el número de veces que la clase fue observada.

Veamos un ejemplo. Supongamos que se tiene una variable cualitativa con posibles valores "Bajo", "Medio" y "Alto". Supongamos además que se tienen 46 observaciones de esta variable con un número de observaciones de cada valor como se muestra en la Figura 2.2.

Clase	Frecuencia
C_1 = Bajo	15
C_2 = Medio	21
C_3 = Alto	10

Tabla 2.2: Tres clases y sus frecuencias.

La información de las 46 observaciones se presenta de manera resumida en la tabla de la Figura 2.2, y es claro que este tipo de representaciones tabulares son muy útiles. Podemos además presentar esta misma información de manera gráfica como en la Figura 2.8.

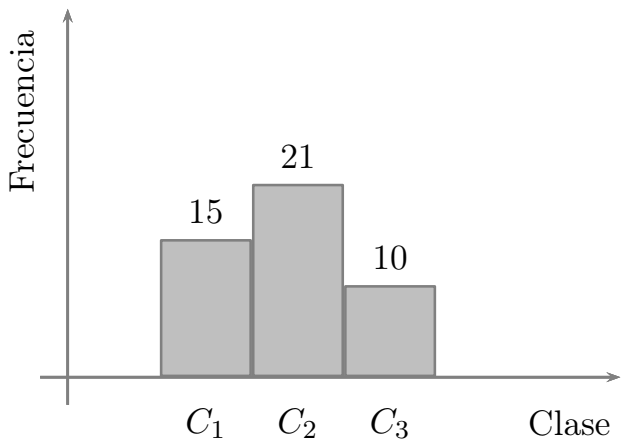


Figura 2.8: Ejemplo de gráfica de frecuencias de clases.

En particular, se trata de una variable cualitativa ordinal pues sus valores poseen un orden. Por ello hemos colocado naturalmente el valor C_1 a la izquierda del valor C_2 y éste a la izquierda del valor C_3. También se pueden hacer este tipo de representaciones gráficas para las frecuencias de los valores de una variable cualitativa nominal, en tal caso no es relevante el orden en el que se grafican las frecuencias.

Veamos ahora un ejemplo de una variable cuantitativa. Supongamos que se registra el número de hijos en 90 familias. Las frecuencias observadas para cada valor de la variable se muestran de manera tabular en la Figura 2.3 y de manera gráfica en la Figura 2.9.

Número de hijos por familia							
Valor	0	1	2	3	4	5	6
Frecuencia	25	40	16	5	1	1	2

Tabla 2.3: Ejemplo de frecuencias de valores.

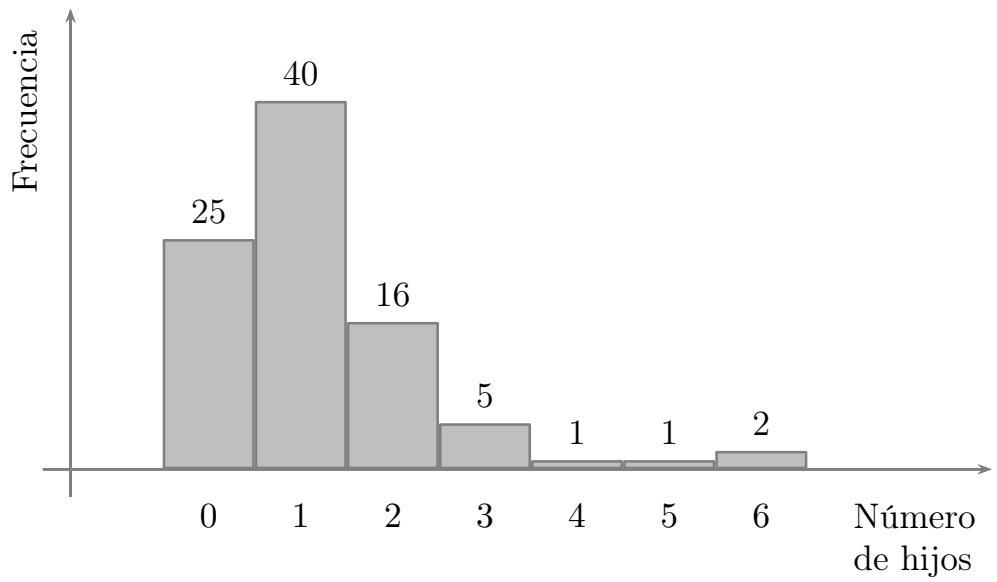


Figura 2.9: Ejemplo de gráfica de frecuencias de valores.

Frecuencias absolutas acumuladas

Cuando las clases $C_1, C_2, \ldots, C_k$ poseen un orden natural y se han definido de menor a mayor como indica el subíndice, es decir, $C_1 \leqslant C_2 \leqslant \cdots \leqslant C_k$, es útil también considerar las frecuencias acumuladas. Definiremos estas frecuencias a continuación.

La **frecuencia acumulada** de una clase (categoría o conjunto de valores) es el número total de veces que la clase considerada, junto con las clases anteriores, fueron observadas.

Es decir, si como antes, $f_1, \ldots, f_k$ denotan las frecuencias de las clases $C_1, \ldots, C_k$, entonces la frecuencia acumulada de la clase C_j es la suma

$$f_1 + f_2 + \cdots + f_j.$$

Los valores de estas frecuencias (absolutas) acumuladas se muestran en la tercera columna de la Tabla 2.4.

Clase	Frecuencia (absoluta)	Frecuencia (absoluta) acumulada
C_1	f_1	f_1
C_2	f_2	$f_1 + f_2$
C_3	f_3	$f_1 + f_2 + f_3$
$\vdots$	$\vdots$	$\vdots$
C_k	f_k	$f_1 + \cdots + f_k$

Tabla 2.4: Ejemplo general de clases y sus frecuencias acumuladas.

Por ejemplo, las 46 observaciones de la variable cualitativa ordinal con valores "Bajo", "Medio" y "Alto" considerada antes, tiene frecuencias y frecuencias acumuladas como se muestra en la Tabla 2.5. La gráfica de las frecuencias acumuladas se muestra en la Figura 2.10.

Clase	Frecuencia	Frecuencia acumulada
C_1 = Bajo	15	15
C_2 = Medio	21	36
C_3 = Alto	10	46

Tabla 2.5: Ejemplo de clases y sus frecuencias acumuladas.

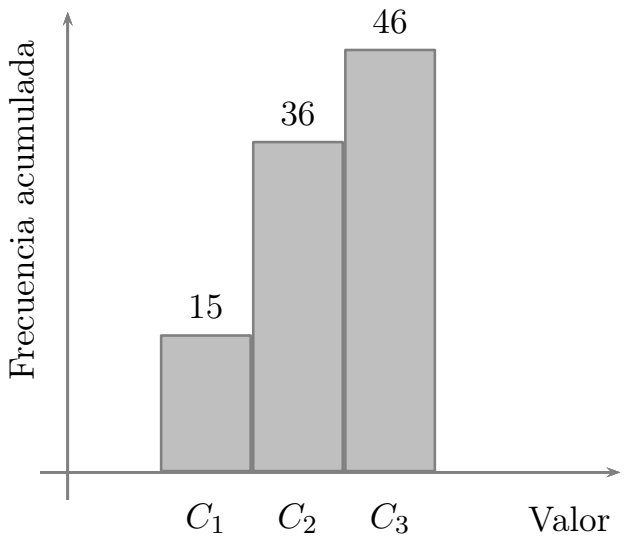


Figura 2.10: Ejemplo de gráfica de frecuencias acumuladas de clases.

Como un segundo ejemplo, consideremos nuevamente la variable cuantitativa definida como el número de hijos en una familia. Las observaciones de 90 familias tienen las frecuencias y frecuencias acumuladas como se muestran en la Tabla 2.6 y de manera gráfica en la Figura 2.11.

Número de hijos en 90 familias							
Valor	0	1	2	3	4	5	6
Frecuencia	25	40	16	5	1	1	2
Frec. acumulada	25	65	81	86	87	88	90

Tabla 2.6: Ejemplo de frecuencias acumuladas.

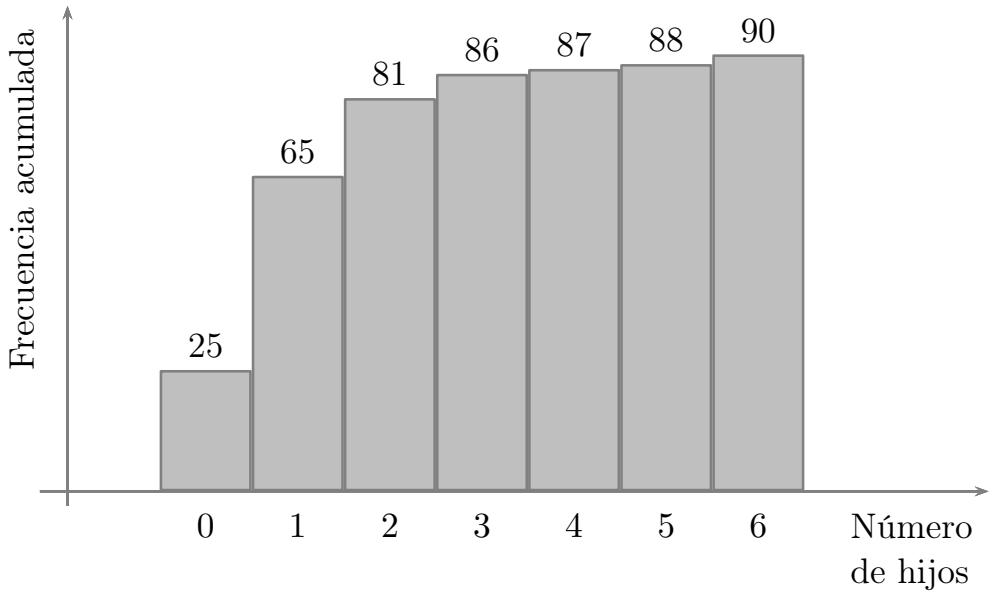


Figura 2.11: Ejemplo de gráfica de frecuencias acumuladas de valores.

Frecuencias relativas

Se pueden definir también las frecuencias relativas al dividir cada frecuencia (absoluta) entre el número total de observaciones. A las cantidades así obtenidas se les llama frecuencias relativas. En este caso no es necesario que haya un orden entre las clases, las frecuencias relativas se pueden también calcular para valores nominales o categóricos.

La **frecuencia relativa** de una clase (categoría o conjunto de valores) es el número de veces que la clase fue observada dividido entre el total de observaciones.

De esta manera, si $f_1, f_2, \ldots, f_k$ son las frecuencias (absolutas), entonces las cantidades $f_1/n, f_2/n, \ldots, f_k/n$ son las frecuencias relativas, suponiendo n observaciones o registros totales. Estas nuevas frecuencias se muestran en la tercera columna de la Tabla 2.7.

Categoría	Frecuencia (absoluta)	Frecuencia relativa	Frecuencia relativa porcentual
C_1	f_1	f_1/n	$100 \cdot f_1/n$
C_2	f_2	f_2/n	$100 \cdot f_2/n$
C_3	f_3	f_3/n	$100 \cdot f_3/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_k	f_k	f_k/n	$100 \cdot f_k/n$
Suma	n	1	100

Tabla 2.7: Forma general de calcular las frecuencias relativas porcentuales.

Observemos que las frecuencias relativas son números en el intervalo $[0, 1]$ y que la suma de todas estas cantidades es 1.

Cuando estas frecuencias relativas se expresan como porcentajes, es decir, cuando se multiplican por 100, se llaman **frecuencias relativas porcentuales**. Estas cantidades son equivalentes a las primeras (sólo que su representación es distinta) y se muestran en la cuarta columna de la tabla de la Figura 2.7.

Por ejemplo, retomemos la variable cualitativa con valores "Bajo", "Medio" y "Alto". En la Tabla 2.8 se recuerdan las frecuencias de cada uno de estos valores con un total de 46 observaciones. En la tercera columna de esta tabla se muestran las frecuencias relativas y en la cuarta columna se muestran las frecuencias relativas expresadas en porcentajes. Observe que debido a la insuficiente precisión en los cálculos numéricos, la suma de las frecuencias relativas no es 1. El mismo fenómeno de falta de precisión produce que la suma de las frecuencias porcentuales no sea 100 %.

Clase	Frecuencia	Frecuencia relativa	Frecuencia relativa porcentual
C_1 = Bajo	15	$15/46 = 0.326$	32.6 %
C_2 = Medio	21	$21/46 = 0.456$	45.6 %
C_3 = Alto	10	$10/46 = 0.217$	21.7 %

Tabla 2.8: Ejemplo del cálculo de frecuencias relativas porcentuales.

Se pueden elaborar gráficas de estas frecuencias relativas y para este ejemplo la gráfica es muy similar a la presentada en la Figura 2.8.

Para el ejemplo, de la variable cuantitativa definida como el número de hijos en una familia, en la Tabla 2.9 se recuerdan las frecuencias observadas de cada valor y se muestran las frecuencias relativas correspondientes. Las gráficas de las frecuencias relativas son similares a la gráfica de la Figura 2.9.

Número de hijos en 90 familias							
Valor	0	1	2	3	4	5	6
Frecuencia	25	40	16	5	1	1	2
Frec. relativa	0.27	0.44	0.17	0.05	0.01	0.01	0.02
Frec. rel. porcentual	27 %	44 %	17 %	5 %	1 %	1 %	2 %

Tabla 2.9: Ejemplo del cálculo de frecuencias relativas porcentuales.

Frecuencias relativas acumuladas

Considerando nuevamente el caso cuando las categorías $C_1, C_2, \ldots, C_k$ poseen un orden natural y se han definido de menor a mayor como indica el subíndice, se pueden definir también las frecuencias relativas acumuladas.

La **frecuencia relativa acumulada** de una clase (categoría o conjunto de valores) es la suma de las frecuencias relativas anteriores e inclusive la clase en cuestión.

Es decir, la frecuencia relativa acumulada de la clase C_j es la suma

$$f_1/n + \cdots + f_j/n.$$

Los valores de estas frecuencias relativas acumuladas se muestran en la tercera columna de la Tabla 2.10.

Clase	Frecuencia relativa	Frecuencia relativa acumulada
C_1	f_1/n	f_1/n
C_2	f_2/n	$f_1/n + f_2/n$
C_3	f_3/n	$f_1/n + f_2/n + f_3/n$
$\vdots$	$\vdots$	$\vdots$
C_k	f_k/n	$f_1/n + \cdots + f_k/n$

Tabla 2.10: Forma general de calcular las frecuencias relativas porcentuales.

Clase	Frecuencia	Porcentaje (%)
C_1	f_1	$\frac{f_1}{n} \cdot 100$
C_2	f_2	$\frac{f_2}{n} \cdot 100$
$\vdots$	$\vdots$	$\vdots$
C_k	f_k	$\frac{f_k}{n} \cdot 100$
Suma	$n = \sum_{i=1}^{k} f_i$	100

Tabla 2.11: Forma general de expresar las frecuencias como porcentajes.

Por ejemplo, para la variable cualitativa con valores "Bajo", "Medio" y "Alto", las frecuencias relativas acumuladas se muestran en la Tabla 2.12.

Clase	Frecuencia	Frecuencia relativa	Frecuencia relativa acumulada
C_1 = Bajo	15	0.326	0.326
C_2 = Medio	21	0.456	0.692
C_3 = Alto	10	0.217	0.902

Tabla 2.12: Ejemplo del cálculo de las frecuencias relativas acumuladas.

Cuantiles

Consideremos nuevamente que $x_1, \ldots, x_n$ es un conjunto de n observaciones de una cierta variable cuantitativa de interés, y que estos valores se ordenan de menor a mayor, conservando las repeticiones. Un cuantil es un número

que separa a los datos en dos partes: un cierto porcentaje de los datos son menores o iguales al cuantil y el porcentaje complementario corresponde a datos que son mayores o iguales al cuantil.

Para dar una definición más precisa de cuantil consideraremos que p es un número cualquiera conocido tal que $0 < p \leqslant 1$. Este valor determinará los porcentajes de los que hablamos en el párrafo anterior. Por ejemplo, podemos suponer que $p = 0.2$. Entonces un cuantil es un número c tal que la proporción de valores x_i que son menores o iguales a c es del 20 % y, al mismo tiempo, la proporción de valores x_i que son mayores o iguales a c es el porcentaje complementario, esto es, el 80 %. En este caso, al número c se le llama cuantil de orden $p = 0.2$ o cuantil al 20 % y se le denota por $c_{0.2}$. Véase la Figura 2.12.

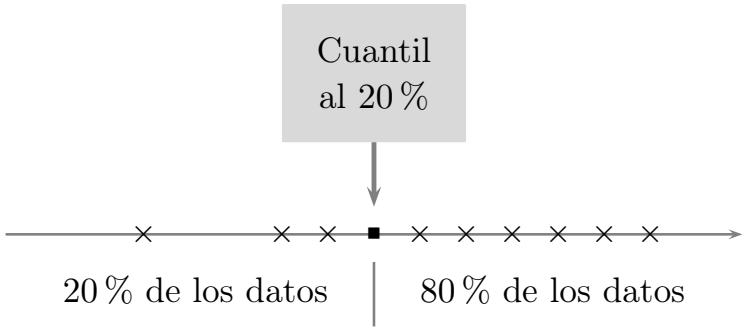


Figura 2.12

En general, podemos tener cuantiles al 5 %, 10 %, 50 %, o cualquier otro porcentaje dado por la expresión $100p\,\%$, con $0 < p \leqslant 1$. Con las ideas introductorias anteriores, podemos ahora dar la definición formal de cuantil para un conjunto de datos numéricos.

Un **cuantil** es un número c tal que cumple las siguientes dos condiciones al mismo tiempo:

$$\frac{\#\{\, x_i \,:\, x_i \leqslant c \,\}}{n} \geqslant p \qquad \text{y} \qquad \frac{\#\{\, x_i \,:\, x_i \geqslant c \,\}}{n} \geqslant 1 - p.$$

Recordemos que si A es un conjunto, entonces la expresión $\#A$ representa la cardinalidad o número de elementos en el conjunto A. De este modo la primera desigualdad que aparece en el recuadro anterior establece que la proporción de observaciones menores o iguales al cuantil c es por lo menos p. La segunda desigualdad establece que la proporción de observaciones que son mayores o iguales a c es por los menos $1 - p$.

Observemos que, a diferencia de lo mencionado antes en la parte introductoria, en la definición formal de cuantil se pide que el porcentaje de datos a la izquierda del cuantil sea por lo menos del $100p\,\%$ (y no necesariamente este porcentaje exacto). Análogamente, el porcentaje de datos a la derecha del cuantil es por lo menos del $100(1-p)\,\%$ (y no necesariamente este porcentaje exacto).

Como hemos mencionado antes, al número c se le llama cuantil p, o cuantil de orden p, o cuantil al $100p\,\%$, y para hacer referencia a la probabilidad p se le denota por $c(p)$, o c_p . En la literatura pueden encontrarse también los símbolos $Q(p)$ o Q_p para denotar al cuantil de orden p. La letra Q proviene del término en inglés *Quantile*.

En ocasiones conviene referirse a los cuantiles que dividen al conjunto de datos en ciertos porcentajes particulares. Por ejemplo, tenemos los siguientes casos.

- Cuando $p = 0.25, 0.50$ ó 0.75, a los cuantiles correspondientes se le llama **cuartiles**, y se usan las expresiones: primer cuartil, segundo cuartil y tercer cuartil, respectivamente.
- Cuando $p = 0.1, 0.2, \ldots, 0.9$, a los cuantiles correspondientes se les llama **deciles**. Podemos referirnos al primer decil de un conjunto de datos, al segundo decil, etcétera.

- En otras ocasiones se requiere dividir al conjunto de datos en cien porcentajes iguales, y entonces cuando $p = 0.01, 0.02, \ldots, 0.99$ a los cuantiles correspondientes se les llama **percentiles**.

El cálculo de los cuantiles puede ser un tanto delicado, así es que daremos varios ejemplos y otros comentarios para fortalecer su entendimiento.

Ejemplo 2.1 Supongamos que tenemos una colección de $n = 2$ números distintos y que éstos son:

$$0, 1.$$

Podemos representar estos dos valores como puntos en un eje horizontal como se muestra en la Figura 2.13. Además, dado que únicamente tenemos dos valores, consideraremos que cada uno de ellos tiene un peso de 0.5 .

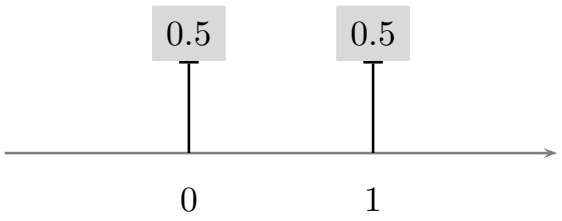


Figura 2.13: Dos valores y sus frecuencias relativas.

Para cada valor de c en el eje horizontal encontraremos los cocientes

$$\frac{\#\{\, x_i \,:\, x_i \leqslant c \,\}}{n} \quad \text{y} \quad \frac{\#\{\, x_i \,:\, x_i \geqslant c \,\}}{n}.$$

Los valores de estos cocientes se encuentran en la segunda y tercera columnas de la Tabla 2.13. Explicaremos a continuación la información que aparece en esta tabla. En la primera columna aparecen los valores de c. Estos valores están separados en cinco casos: primero los valores en el intervalo $(-\infty, 0)$, luego el valor 0, luego los valores en el intervalo $(0, 1)$, luego el valor 1 y finalmente los valores en el intervalo $(1, \infty)$. En la segunda y tercera columnas aparece el cálculo de los cocientes indicados en el encabezado considerando un valor de c según los cinco casos mencionados. Por ejemplo, para cualquier

valor de c en el intervalo $(-\infty, 0)$, el conjunto $\{x_i : x_i \leqslant c\}$ es vacío y por lo tanto el cociente es 0. En cambio, el conjunto $\{x_i : x_i \geqslant c\}$ es la totalidad de los datos y el cociente es 1.

c	$\#\{x_i : x_i \leqslant c\}/n$	$\#\{x_i : x_i \geqslant c\}/n$
$(-\infty, 0)$	0	1
0	0.5	1
$(0, 1)$	0.5	0.5
1	1	0.5
$(1, \infty)$	1	0

Tabla 2.13: Ejemplo del cálculo de frecuencias acumuladas relativas para obtener los cuantiles.

En la tabla aparece separado el valor $c = 0$. Veamos este caso. Tenemos que $\{x_i : x_i \leqslant 0\} = \{0\}$ y $\{x_i : x_i \geqslant 0\} = \{0, 1\}$, y es así como se obtienen los valores 0.5 y 1 que aparecen en este renglón de la tabla. De manera análoga se obtienen los siguientes renglones de esta tabla. El lector debe asegurarse de poder comprobar los valores de esta tabla apoyándose en el diagrama de la Figura 2.13. Ahora, ¿para qué hemos hecho esto? Pues resulta que la tabla de la Figura 2.13 puede proporcionarnos el valor de cualquier cuantil de este conjunto pequeño de datos. Veamos algunos ejemplos.

a) El cuantil al 10 % es el valor $c = 0$ pues se cumple

$$\begin{aligned} \frac{\#\{x_i : x_i \leqslant 0\}}{n} &\geqslant 0.10, \\ \frac{\#\{x_i : x_i \geqslant 0\}}{n} &\geqslant 0.90. \end{aligned}$$

De esta manera el valor 0 divide al conjunto de datos en dos partes, a la izquierda de este valor queda (por lo menos) el 10 % y a la derecha queda

(por lo menos) el 90 %.

b) Cualquier valor de c en el intervalo $(0, 1)$ es un cuantil al 50 % pues para cualquier valor en este intervalo se tiene el cumplimiento de las siguientes condiciones

$$\begin{aligned} \frac{\#\{\, x_i \,:\, x_i \leqslant c \,\}}{n} &\geqslant 0.50, \\ \frac{\#\{\, x_i \,:\, x_i \geqslant c \,\}}{n} &\geqslant 0.50. \end{aligned}$$

Cualquier valor de c en el intervalo $(0, 1)$ divide al conjunto de datos en dos partes: a la izquierda de c queda el 50 % de los datos y a la derecha el otro 50 %. El punto medio del intervalo $(0, 1)$ es el valor 0.5. Tomaremos este valor como el cuantil al 50 %.

c) El cuantil al 70 % es el valor $c = 1$ pues se cumple

$$\begin{aligned} \frac{\#\{\, x_i \,:\, x_i \leqslant 1 \,\}}{n} &\geqslant 0.70, \\ \frac{\#\{\, x_i \,:\, x_i \geqslant 1 \,\}}{n} &\geqslant 0.30. \end{aligned}$$

Por lo tanto, el valor 1 divide al conjunto de datos en dos partes: a la izquierda de este valor queda (por lo menos) el 70 % de los datos y a la derecha queda (por lo menos) el 30 %.

Los tres cálculos anteriores pueden reproducirse a un mismo tiempo en el paquete estadístico `R` mediante el siguiente código.

R

```
> x <- c(0,1)
> quantile(x,c(0.1,0.5,0.7),type=2)
[1] 10% 50% 70%
0.0 0.5 1.0
```

•

Existen varias formas en las que puede definirse un cuantil cuando todos los valores dentro de un intervalo satisfacen las condiciones de la definición. En el presente trabajo hemos adoptado la convención de que, en tales casos, el punto medio del intervalo es el cuantil correspondiente. Esta convención es compatible con la definición de mediana, o cuantil al 50 %, mencionada antes. La especificación `type=2` dentro del comando `quantile()`, usado en el recuadro anterior, indica esta forma de calcular los cuantiles. Veamos otros ejemplos del cálculo de cuantiles con otro conjunto de datos.

Ejemplo 2.2 Supongamos que tenemos ahora una colección de $n = 4$ números y que éstos son:

$$0, 0, 1, 2.$$

Esta vez no todos son distintos. Nuevamente representemos estos datos como puntos en un eje horizontal como se muestra en la Figura 2.14. Como el valor 0 aparece dos veces tiene un peso igual a $2/4 = 1/2 = 0.5$. Los otros dos puntos tienen peso $1/4 = 0.25$.

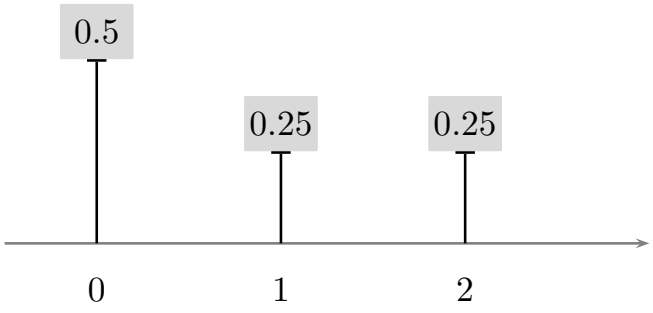


Figura 2.14: Tres valores y sus frecuencias relativas.

Para cada valor de c en el eje horizontal, encontraremos los cocientes

$$\frac{\#\{\, x_i \,:\, x_i \leqslant c \,\}}{n} \quad \text{y} \quad \frac{\#\{\, x_i \,:\, x_i \geqslant c \,\}}{n}.$$

Los valores de estos cocientes se encuentran en la segunda y tercera columnas de la Tabla 2.14, la cual explicaremos a continuación. En la primera

columna aparecen los valores de c. Estos valores están separados ahora en siete casos: primero los valores en el intervalo $(-\infty, 0)$, luego el valor 0, luego los valores en el intervalo $(0, 1)$, luego el valor 1, luego los valores en el intervalo $(1, 2)$, luego el valor 2 y finalmente los valores en el intervalo $(2, \infty)$. En la segunda y tercera columnas aparece el cálculo de los cocientes indicados en el encabezado considerando un valor de c según los siete casos mencionados. Por ejemplo, para cualquier valor de c en el intervalo $(0, 1)$, el conjunto $\{x_i : x_i \leqslant c\}$ es igual a $\{0\}$ y por lo tanto el cociente es 0.5. En contraparte, el conjunto $\{x_i : x_i \geqslant c\}$ es igual a $\{1, 2\}$ y el cociente es 0.5.

c	$\#\{\, x_i \,:\, x_i \leqslant c \,\}/n$	$\#\{\, x_i \,:\, x_i \geqslant c \,\}/n$
$(-\infty, 0)$	0	1
0	0.5	1
$(0, 1)$	0.5	0.5
1	0.75	0.5
$(1, 2)$	0.75	0.25
2	1	0.25
$(2, \infty)$	1	0

Tabla 2.14: Ejemplo del cálculo de frecuencias acumuladas relativas para obtener los cuantiles.

De manera análoga se obtienen los otros renglones de esta tabla. Es posible comprobar los valores de esta tabla a partir del diagrama de la Figura 2.14. Esta tabla puede proporcionarnos el valor de cualquier cuantil de este conjunto de cuatro datos. Veamos algunos ejemplos.

a) El cuantil al 20 % es el valor $c = 0$ pues se cumple

$$\begin{aligned} \frac{\#\{\, x_i \,:\, x_i \leqslant 0 \,\}}{n} &\geqslant 0.20, \\ \frac{\#\{\, x_i \,:\, x_i \geqslant 0 \,\}}{n} &\geqslant 0.80. \end{aligned}$$

El valor 1 divide al conjunto de datos en dos partes: a la izquierda de este valor queda (por lo menos) el 20 % de los datos y a la derecha queda (por lo menos) el 80 %.

b) Cualquier valor de c en el intervalo $(0, 1)$ es un cuantil al 50 % pues para cualquier valor de c en el intervalo indicado tenemos el cumplimiento de las siguientes condiciones

$$\begin{aligned} \frac{\#\{\, x_i \,:\, x_i \leqslant c \,\}}{n} &\geqslant 0.50, \\ \frac{\#\{\, x_i \,:\, x_i \geqslant c \,\}}{n} &\geqslant 0.50. \end{aligned}$$

Cualquier valor de c en el intervalo $(0, 1)$ divide al conjunto de datos en dos partes: a la izquierda de c queda el 50 % de los datos y a la derecha el otro 50 %. El punto medio del intervalo $(0, 1)$ es el valor 0.5. Tomaremos este valor como el cuantil al 50 %.

c) El cuantil al 80 % es el valor $c = 2$ pues se cumple

$$\begin{aligned} \frac{\#\{\, x_i \,:\, x_i \leqslant 2 \,\}}{n} &\geqslant 0.80, \\ \frac{\#\{\, x_i \,:\, x_i \geqslant 2 \,\}}{n} &\geqslant 0.20. \end{aligned}$$

Por lo tanto, el valor 2 divide al conjunto de datos en dos partes: a la izquierda de este valor queda (por lo menos) el 80 % de los datos y a la derecha queda (por lo menos) el 20 %.

En el paquete estadístico `R`, los tres cálculos anteriores pueden reproducirse a un mismo tiempo usando el siguiente código.

R

```
> x <- c(0,0,1,2)
> quantile(x,c(0.2,0.5,0.8),type=2)
[1] 20% 50% 80%
0.0 0.5 2.0
```

En la sección que trata sobre la función de distribución empírica y que inicia en el página 117, veremos una forma gráfica alternativa para calcular los cuantiles de un conjunto de datos numéricos.

Coeficiente de asimetría (*Skewness*)

Un conjunto de datos numéricos es simétrico si estas cantidades se encuentran distribuidas simétricamente alrededor de la media. Véase la Figura 2.15 en donde se muestra un ejemplo gráfico de datos simétricos y otro ejemplo de datos no simétricos.

Datos simétricos Datos no simétricos

Figura 2.15: Ejemplos gráficos de simetría y no simetría.

La cantidad que llamaremos coeficiente de asimetría (en inglés *skewness*) es una medida de la asimetría (falta de simetría) de un conjunto de datos numéricos $x_1, \ldots, x_n$. Si $\bar{x}$ es la media y s es la desviación estándar, entonces el coeficiente de asimetría se define como el siguiente número.

$$\text{sk} = \frac{1}{s^3}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3\right)$$

Recordemos que s^2 denota la varianza, en consecuencia, el término s^3 se calcula de la forma siguiente

$$s^3 = (s^2)^{3/2} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}.$$

El coeficiente de asimetría no posee unidad de medición, es un número que puede ser positivo, negativo o cero. Su signo es positivo cuando la gráfica de frecuencias de los datos presenta una cola más alargada hacia la derecha de la media. Este tipo de comportamiento general se muestra en la gráfica derecha de la Figura 2.16 y es un indicativo de que existen datos a la derecha y alejados de la media de tal forma que las cantidades $(x_i - \bar{x})^3$ son comparativamente grandes y con signo positivo.

En cambio, el signo del coeficiente de asimetría es negativo cuando la gráfica de frecuencias presenta una cola más alargada hacia la izquierda de la media. Este comportamiento se muestra en la parte izquierda de la Figura 2.16. En este caso existen datos a la izquierda y alejados de la media de tal forma que las cantidades $(x_i - \bar{x})^3$ son grandes y con signo negativo. Por supuesto, se puede tener una gráfica de frecuencias sin presentar con claridad ninguno de estos dos tipos de comportamientos, pero el coeficiente de asimetría proporciona una cuantificación acerca de la tendencia global de los datos hacia alguno de estos dos posibles escenarios.

Asimetría negativa Asimetría positiva

Figura 2.16: Ejemplos gráficos de asimetría negativa y positiva.

Veamos ahora el caso simétrico. Verificaremos que en esta situación el coeficiente de asimetría se hace cero. La simetría de los datos significa que por cada dato x_i a la izquierda de $\bar{x}$ hay otro dato a la derecha y a la misma distancia de este punto central. Al considerar las cantidades $(x_i - \bar{x})^3$, los dos datos producirán la misma cantidad, pues ambos se encuentran a la misma distancia de $\bar{x}$, pero tendrán signos contrarios, y al hacer la suma estas cantidades se cancelan una con otra. De esta manera la suma total es cero.

Es importante advertir al lector que existen otras formas de definir un coeficiente de asimetría para un conjunto de datos o una distribución. A la definición que hemos visto se le conoce como coeficiente de asimetría de Fisher-Pearson, pero existen otras pequeñas variaciones en la definición y es prudente consultar el manual del paquete de cómputo utilizado, si es el caso, para verificar la forma exacta de su cálculo. Por ejemplo, el paquete `R` calcula el coeficiente de asimetría como lo hemos estudiado y se puede calcular usando el código que aparece abajo. Para que estas instrucciones funcionen se debe instalar previamente el paquete `moments`.

R

```
> library(moments)
> x <- c(0,0,0,1,1,2)
> skewness(x)
[1] 0.626099
```

En términos de los momentos centrales m_2 y m_3, el coeficiente de asimetría

se puede escribir de la siguiente forma.

$$sk = \frac{m_3}{m_2^{3/2}}.$$

Curtosis

Sea $x_1, \ldots, x_n$ una colección de datos numéricos con media $\bar{x}$ y desviación estándar s. La curtosis, que denotaremos por la letra k, es un número que se define de la siguiente manera.

$$\mathrm{k} = \frac{1}{s^4}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4\right)$$

Recordemos que s^2 denota la varianza, en consecuencia, el término s^4 denota la varianza al cuadrado y se calcula de la forma siguiente

$$s^4 = (s^2)^2 = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2.$$

La curtosis es un número mayor o igual a cero que no tiene una unidad de medición. Dado que las cantidades $(x_i - \bar{x})^4$ aparecen involucradas en el cálculo de la curtosis, cuando una observación x_i dista mucho de la media $\bar{x}$, al elevar esta distancia a la cuarta potencia hace que se magnifiquen las distancias grandes. Por lo tanto, una curtosis grande puede indicar un mayor número de datos alejados de la media, hacia uno u otro lado, y por lo tanto a la curtosis se le interpreta como una medida de la forma de las colas de la distribución o del conjunto de datos.

Por la expresión "forma de las colas" nos referimos aquí a si éstas son amplias o bien ligeras (o inexistentes). Si son de una forma o de otra, esto afecta la forma de un posible pico que presente la frecuencia de los datos y de allí surgen interpretaciones de la curtosis como una medida del tipo de pico de los datos. Estas interpretaciones están sujetas a debate y por ahora no existe una interpretación aceptada de manera general.

Es claro que en términos de los momentos centrales, la curtosis puede escribirse de la siguiente manera $k = m_4/m_2^2$. En el paquete `R` se puede calcular la curtosis de un conjunto de datos mediante el comando `kurtosis()`, como se muestra en el siguiente recuadro.

R

```
> library(moments)
> x <- c(0,0,0,1,1,2)
> kurtosis(x)
[1] 2.04
```

Advertencia: se usa también con el nombre de curtosis (o bien *excess kurtosis*) la cantidad que aparece abajo. Debido a que la curtosis de la distribución normal estándar es igual a 3, con esta nueva definición la curtosis de la distribución normal es ahora cero.

$$k_3 = \frac{1}{s^4}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4\right) - 3.$$

De esta manera se toma el tipo de cola de la distribución normal como punto de referencia y se adoptan los siguientes nombres:

- Leptocúrtica ($k_3 > 0$): Decaimiento rápido, colas ligeras. Véase la Figura 2.17 (a).

- Mesocúrtica ($k_3 = 0$): Curva normal. Véase la Figura 2.17 (b).

- Platicúrtica ($k_3 < 0$): Decaimiento lento, colas amplias. Véase la Figura 2.17 (c).

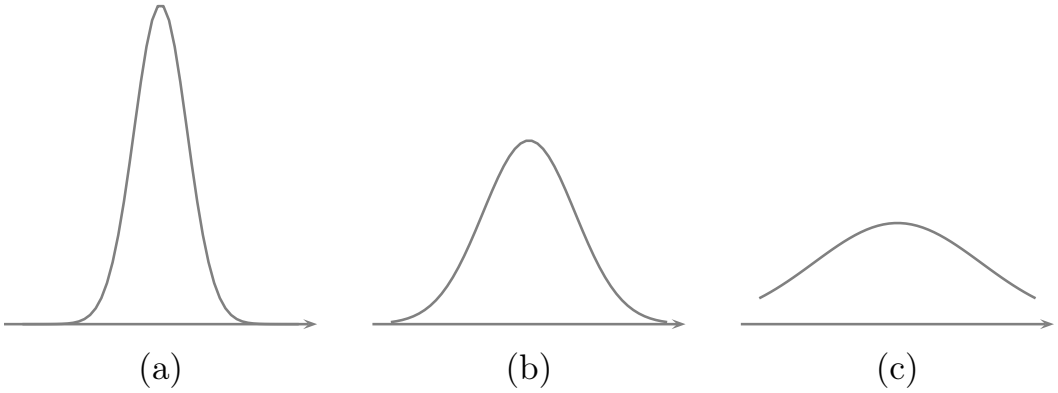


Figura 2.17: (a) Curva leptocúrtica, (b) mesocúrtica (normal) y (c) platicúrtica.

Debe considerarse que los tres tipos de comportamientos indicados son de tipo general y que, dependiendo del signo de la curtosis k_3, es que esta cantidad puede sugerir una tendencia de los datos hacia uno u otro tipo de comportamiento. Es claro que un conjunto de datos no necesariamente presenta uno de estos tres tipos de forma en su gráfica de frecuencias. El valor de la curtosis es únicamente una insinuación hacia alguno de estos tres tipos de comportamientos.

Descripciones numéricas para datos agrupados

En ocasiones la información disponible para una variable cuantitativa se encuentra agrupada en categorías o subconjuntos de valores de la variable. Más específicamente, supongamos que en lugar de tener las observaciones o registros individuales $x_1, \ldots, x_n$, tenemos agrupamientos de valores $C_1, \ldots, C_k$ junto con las frecuencias $f_1, \ldots, f_k$ que indican el número de veces que se observó cada agrupamiento. Por ejemplo, para la variable cuantitativa definida como el número de años cumplidos de una persona, podemos tener la siguiente información agrupada:

Categoría	Frecuencia
$C_1 = \{10, 11, 12\}$	$f_1 = 4$
$C_2 = \{13, 14, 15\}$	$f_2 = 3$
$C_3 = \{16, 17, 18\}$	$f_3 = 2$
$C_4 = \{19, 20, 21\}$	$f_4 = 3$

Esto significa que hay 4 personas con edad entre 10 y 12 años, hay 3 personas con edad entre 13 y 15 años, etcétera.

El problema es el siguiente: ¿cómo podemos calcular las descripciones numéricas como la media y la varianza en este caso? Existen dos perspectivas que explicaremos a continuación.

1. **Primera aproximación.** Se determina una marca de clase para cada categoría y se considera que la marca de clase se observó tantas veces como indica la frecuencia de la categoría. De esta manera se construyen observaciones individuales aproximadas y se pueden aplicar ahora todas las definiciones y fórmulas vistas antes. Para el ejemplo anterior, tenemos la siguiente tabla:

Categoría	Frecuencia	Marca de clase	Observaciones (aproximadas)
$C_1 = \{10, 11, 12\}$	$f_1 = 4$	11	$11, 11, 11, 11$
$C_2 = \{13, 14, 15\}$	$f_2 = 3$	14	$14, 14, 14$
$C_3 = \{16, 17, 18\}$	$f_3 = 2$	17	$17, 17$
$C_4 = \{19, 20, 21\}$	$f_4 = 3$	20	$20, 20, 20$

   En este ejemplo particular es fácil determinar una marca de clase: se ha escogido simplemente el valor central de cada categoría. En general,

alguna explicación razonable debe utilizarse para justificar la elección de una marca de clase para cada categoría.

2. **Segunda aproximación.** Se escogen tantos valores numéricos dentro de una categoría como indica la frecuencia. Por ejemplo, pueden escogerse valores equiespaciados si esto es posible. Como antes, se procede a aplicar las fórmulas a la colección de valores numéricos así generados. Para el ejemplo anterior, podemos hacer la siguiente selección de valores:

Categoría	Frecuencia	Valores seleccionados
$C_1 = \{10, 11, 12\}$	$f_1 = 4$	$10, 11, 11, 12$
$C_2 = \{13, 14, 15\}$	$f_2 = 3$	$13, 14, 15$
$C_3 = \{16, 17, 18\}$	$f_3 = 2$	$16, 18$
$C_4 = \{19, 20, 21\}$	$f_4 = 3$	$19, 20, 21$

En este caso también es conveniente justificar la elección de los valores dentro de una categoría.

Debe enfatizarse que, en cualquiera de las dos perspectivas explicadas, la información producida es únicamente una aproximación, pues se ha perdido información al considerar el agrupamiento de valores.

RESUMEN DE FÓRMULAS

Descripciones numéricas de un conjunto de datos $x_1, \ldots, x_n$

Media	$\bar{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$
Moda	Dato con mayor frecuencia.
Mediana	Dato ordenado de en medio.
Varianza	$s^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2.$
Desviación estándar	$s = \sqrt{\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2}.$
Desviación media	$dm = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert.$
Rango	$r = x_{(n)} - x_{(1)}.$
Coeficiente de variación	$cv = \dfrac{s}{\bar{x}}.$
Momentos	$m'_k = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i^k.$
Momentos centrales	$m_k = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^k.$
Cuantil al $100p\,\%$	Al menos el $100p\,\%$ de los datos son menores al cuantil y al menos $100(1-p)\,\%$ de los datos son mayores al cuantil.
Asimetría	$\mathrm{sk} = \dfrac{1}{s^3} \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^3 \right)$
Curtosis	$\mathrm{k} = \dfrac{1}{s^4} \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^4 \right)$

Ejercicios

Media

11. Calcule la media del siguiente conjunto de datos.

    a) $-1, 5, 0, 3, 3, 2, 1, 0.$

    b) $2.5, 3.2, 2.7, 2.0, 3.4, 3.7, 5.2, 2.0, 2.5, 4.0, 2.3.$

    c) $20, 32, 25, 27, 30, 28, 31, 20, 30.$

    d) $0, 1, -1, 0, 1, 1, 0, 0, 1.$

12. Calcule la media del siguiente conjunto de datos.

    a) $12, 15, 30, 23, 10.$

    b) $23, 10, 15, 30, 12.$

    c) $30, 15, 12, 10, 23.$

    d) $15, 30, 23, 10, 12.$

13. Considere el conjunto de cinco datos como aparece abajo. Determine el valor del dato faltante x_5 si la media es 6.

$$\begin{aligned} x_1 &= 3, \\ x_2 &= 8, \\ x_3 &= 4, \\ x_4 &= 7, \\ x_5 &= ? \end{aligned}$$

14. Calcule la media de los siguientes conjuntos de datos. La primera columna corresponde al conjunto de datos original. El resto de las columnas se obtienen según la operación indicada en el encabezado.

x	$x+2$	$x-4$	$2x$	$10x$
4	6	0	8	40
7	9	3	14	70
3	5	-1	6	30
2	4	-2	4	20

15. Calcule la media de los datos que aparecen en la tabla de abajo. Además de los valores observados se proporciona el número de veces que cada valor ha sido observado.

Valor	-2	-1	0	1	2
Frecuencia	7	2	1	4	3

16. Los siguientes datos corresponden a las calificaciones aprobatorias de un estudiante y el número de veces que ha obtenido cada calificación en cursos anteriores. Calcule el promedio del estudiante.

Calificación	6	7	8	9	10
Frecuencia	0	2	3	5	3

17. Calcule la media de los dos conjuntos de datos que aparecen en la Figura 2.18. Considere que se tienen tantos datos como puntos aparezcan sobre el valor numérico.

Figura 2.18

18. Indique el punto que corresponde a la media en cada uno de los dos conjuntos de datos que aparecen en la Figura 2.19. Considere que cada punto se observa con la frecuencia indicada arriba de la barra correspondiente.

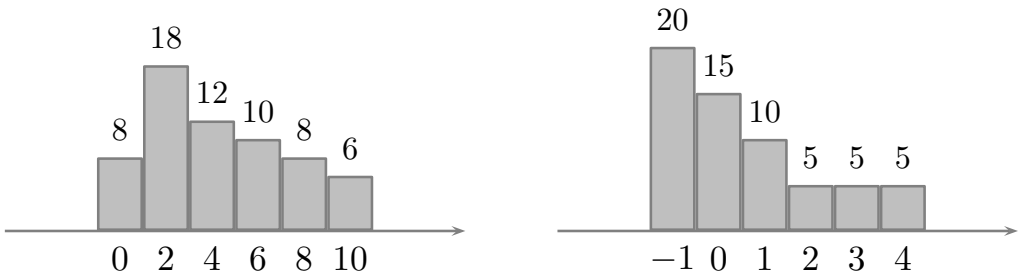


Figura 2.19

19. ¿Cuál es la media de un conjunto que consta de

    a) un único dato?

    b) dos datos idénticos?

    c) dos datos distintos?

    d) mil datos idénticos?

20. Diga falso o verdadero.

    a) La media puede ser cero.

    b) Dos conjuntos de datos distintos pueden tener la misma media.

c) Si se añade un cero a un conjunto de datos la media no cambia.

d) Si se añade un valor positivo a un conjunto de datos la media aumenta.

e) Si se añade un valor negativo a un conjunto de datos la media disminuye.

21. Suponga que los datos numéricos $x_1, \ldots, x_n$ se transforman en los datos $ax_1 + c, \ldots, ax_n + c$, en donde a y c son dos constantes. Sean $y_1, \ldots, y_n$ los nuevos datos y sea $\bar{y}$ su media. Compruebe que

$$\bar{y} = a\bar{x} + c.$$

22. Sea $\bar{x}_n$ la media del conjunto de datos $x_1, \ldots, x_n$. Suponga que se añade un dato adicional x_{n+1} a esta lista y sea $\bar{x}_{n+1}$ la nueva media. ¿Cómo se puede expresar $\bar{x}_{n+1}$ en términos de $\bar{x}_n$ y x_{n+1}? Compruebe que

$$\bar{x}_{n+1} = \frac{1}{n+1}\,(n\,\bar{x}_n + x_{n+1}).$$

23. A un conjunto de datos se le añade un dato adicional y resulta que la media no cambia. ¿Cuál es el dato que se añadió?

24. Sea $\bar{x}_n$ la media del conjunto de datos $x_1, \ldots, x_n$. Suponga que se omite el dato x_i de esta lista y sea $\bar{x}_{n-1}$ la nueva media. ¿Cómo se puede expresar $\bar{x}_{n-1}$ en términos de $\bar{x}_n$ y x_i? Compruebe que

$$\bar{x}_{n-1} = \frac{1}{n-1}\,(n\,\bar{x}_n - x_i).$$

25. Sea $\bar{x}$ la media de $x_1, \ldots, x_n$ y sea $\bar{y}$ la media de $y_1, \ldots, y_m$. ¿Cuál es la media de los datos conjuntos $x_1, \ldots, x_n, y_1, \ldots, y_m$? Compruebe que esta media es

$$\frac{n}{n+m}\,\bar{x} + \frac{m}{n+m}\,\bar{y}.$$

26. Durante su primer año en la universidad un estudiante ha tomado 9 cursos y tiene un promedio de 8.5. Si durante el primer semestre tomó 5 cursos y obtuvo un promedio igual a 8, ¿cuál es el promedio de calificaciones de los cursos tomados en el segundo semestre?

27. Sean $x_1, \ldots, x_n$ observaciones numéricas de una cierta variable de interés. Compruebe que

    a) $\displaystyle\sum_{i=1}^{n} (x_i - \bar{x}) = 0.$

    b) $\displaystyle\sum_{i=1}^{n} (x_i - \bar{x})^2 = (\sum_{i=1}^{n} x_i^2) - n\bar{x}^2.$

28. Diga falso o verdadero: si se le añaden ceros a un conjunto de datos, la media no cambia.

29. Sea $x_1, \ldots, x_n$ un conjunto de datos numéricos con media $\bar{x} \neq 0$. Defina $y_i = x_i/\bar{x}$ para $i = 1, 2, \ldots, n$. Compruebe que $\bar{y} = 1$.

30. *Media geométrica.* Para un conjunto de datos numéricos $x_1, \ldots, x_n$, en donde cada uno de ellos es estrictamente positivo, se define la media geométrica como la raíz n-ésima del producto de todos estos números, es decir,

$$g(x) = \sqrt[n]{x_1 \cdots x_n}$$

    Compruebe las siguientes fórmulas:

    a) $\log g(x) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \log x_i.$

    b) Si $a > 0$ es una constante y ax denota el conjunto de datos en donde cada valor x_i es multiplicado por la constante a, entonces

$$g(ax) = a \cdot g(x).$$

    c) Si y denota otra colección de la misma cantidad de números $y_1, \ldots, y_n$, todos ellos estrictamente positivos, y x/y denota la colección $x_1/y_1, \ldots, x_n/y_n$, entonces

$$g(x/y) = \frac{g(x)}{g(y)}.$$

    Dada su mayor dificultad no se pide hacer aquí, pero puede comprobarse que la media geométrica es siempre menor o igual a la media aritmética, es decir, $g(x) \leqslant \bar{x}$, o más explícitamente,

$$\sqrt[n]{x_1 \cdots x_n} \leqslant \frac{x_1 + \cdots + x_n}{n}.$$

31. *Media armónica.* Para un conjunto de datos numéricos $x_1, \ldots, x_n$, en donde cada uno de ellos es distinto de cero, se define la media armónica como el número

$$h(x) = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}.$$

Suponiendo que $1/x$ denota la colección de datos $1/x_1, \ldots, 1/x_n$, compruebe las siguientes fórmulas:

a) $h(1/x) = \dfrac{1}{\bar{x}}$.

b) $h(x) = \dfrac{n \cdot (x_1 \cdots x_n)}{\sum_{i=1}^{n} (x_1 \cdots x_n)/x_i}$.

Dada su mayor dificultad no se pide hacer aquí, pero puede comprobarse que la media armónica es siempre menor o igual a la media geométrica, y por lo mencionado en el ejercicio anterior, esta última es menor o igual a la media aritmética, es decir, $h(x) \leqslant g(x) \leqslant \bar{x}$, o más explícitamente,

$$\frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \leqslant \sqrt[n]{x_1 \cdots x_n} \leqslant \frac{x_1 + \cdots + x_n}{n}.$$

Moda

32. Suponga que a, b, c, d son los valores de una variable cualitativa. Encuentre las posibles modas del siguiente conjunto de datos.

a) $a, a, a, a, b, b, b, c, c, c, c, d, d, d$.

b) $a, b, c, a, b, c, a, b, c, a, b, c, a, b, c$.

c) $a, b, b, b, b, b, c, d, d, d, d, d$.

d) a, b, c, d.

33. Calcule la moda de los siguientes conjuntos de datos. El primer renglón corresponde al conjunto de datos original. El resto de los renglones se obtienen según la operación indicada.

x	3	5	4	3	5	1	4	3	2	1	2	3
$x+3$	6	8	7	6	8	4	7	6	5	4	5	6
$x/2$	3/2	5/2	2	3/2	5/2	1/2	2	3/2	1	1/2	1	3/2
$x-2$	1	3	2	1	3	-1	2	1	0	-1	0	1
$2x$	6	10	8	6	10	2	8	6	4	2	4	6
$10x$	30	50	40	30	50	10	40	30	20	10	20	30

34. Diga falso o verdadero: si se le añaden ceros a un conjunto de datos, la moda, si existe, no cambia.

Mediana

35. Calcule la mediana del siguiente conjunto de datos.

    a) $-1, 5, 0, 3, 3, 2, 1, 0$.

    b) $2.5, 3.2, 2.7, 2.0, 3.4, 3.7, 5.2, 2.0, 2.5, 4.0, 2.3$.

    c) $20, 32, 25, 27, 30, 28, 31, 20, 30$.

    d) $0, 1, -1, 0, 1, 1, 0, 0, 1$.

36. Calcule la mediana del siguiente conjunto de datos.

    a) $12, 15, 30, 23, 10$.

    b) $23, 10, 15, 30, 12$.

    c) $30, 15, 12, 10, 23$.

    d) $15, 30, 23, 10, 12$.

37. Calcule la mediana de los siguientes conjuntos de datos. La primera columna corresponde al conjunto de datos original. El resto de las columnas se obtienen según la operación indicada en el encabezado.

x	$x+2$	$x+4$	$x-3$	$x/2$	$5x$
4	6	8	1	2	20
7	9	11	4	7/2	35
3	5	7	0	3/2	15
2	4	6	-1	1	10

38. ¿Cuál es la mediana de un conjunto que consta de

    a) un único dato?

    b) dos datos idénticos?

    c) dos datos distintos?

    d) tres datos idénticos?

    e) tres datos distintos?

    f) mil datos idénticos?

39. A un conjunto de datos se le añade un dato adicional y resulta que la mediana no cambia. ¿Cuál es el dato que se añadió?

40. Considere un conjunto de datos cuya mediana es $\tilde{x}$. Diga falso o verdadero.

    a) Si se añade un dato a la izquierda de $\tilde{x}$ y otro dato a la derecha de $\tilde{x}$, entonces la mediana no cambia.

    b) Si se omite un dato a la izquierda de $\tilde{x}$ y se omite otro dato a la derecha de $\tilde{x}$, entonces la mediana no cambia.

41. Calcule la mediana de los dos conjuntos de datos que aparecen en la Figura 2.20. Considere que se tienen tantos datos como puntos aparezcan sobre el valor numérico.

Figura 2.20

42. Indique el punto que corresponde a la mediana en cada uno de los dos conjuntos de datos que aparecen en la Figura 2.21. Considere que cada punto se observa con la frecuencia indicada arriba de la barra correspondiente.

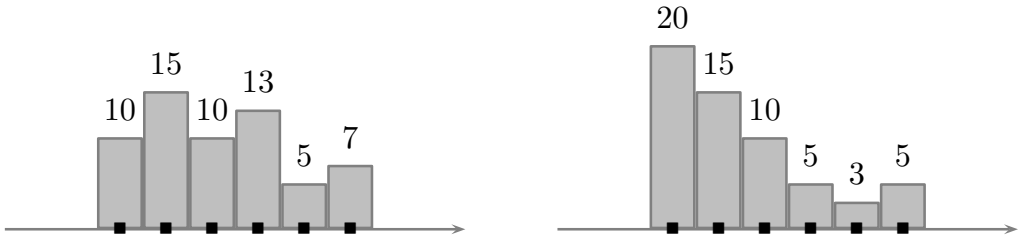


Figura 2.21

43. Diga falso o verdadero: si se le añaden ceros a un conjunto de datos, la mediana no cambia.

44. Considere el conjunto de cinco números: $3, 5, 1, 8, 4$. Suponga que el valor 4 se cambia por el valor 8. Diga si cada una de las siguientes cantidades aumenta, disminuye o no cambia.

    a) La media.

    b) La moda.

    c) La mediana.

Varianza

45. Calcule la varianza del siguiente conjunto de datos.

    a) $-1, 0, 1$.

    b) $2, 2, 2, 2$.

    c) $0, 1, 2$.

    d) $-1, -1, 0, 1, 1$.

46. Considere el conjunto de datos de dos números: x_1 y x_2. Encuentre los valores de estos números si la media es 0 y la varianza es 1.

47. Sean $x_1, \ldots, x_n$ observaciones numéricas de una cierta variable de interés y sea s^2 su varianza. Sea c cualquier constante. Compruebe que

    a) $s^2 = (\frac{1}{n}\sum_{i=1}^{n} x_i^2) - \bar{x}^2$.

    b) $s^2 = [\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2] - (\bar{x} - c)^2$.

48. Sea s_x^2 la varianza del conjunto de datos numéricos $x_1, \ldots, x_n$. Suponga que estos números se transforman en los datos $ax_1 + c, \ldots, ax_n + c$, en donde a y c son dos constantes. Sean $y_1, \ldots, y_n$ los nuevos datos y sea s_y^2 su varianza. Compruebe que

$$s_y^2 = a^2 \cdot s_x^2.$$

49. Calcule la varianza de los siguientes conjuntos de datos. La primera columna corresponde al conjunto de datos original. El resto de las columnas se obtiene según la operación indicada en el encabezado.

x	$x+1$	$x-2$	$2x$	$x/2$	$2x+1$
2	3	0	4	1	5
0	1	-2	0	0	1
0	1	-2	0	0	1
1	2	-1	2	1/2	3
4	5	2	8	2	9

50. Sean $x_1, \ldots, x_n$ observaciones numéricas de una cierta variable de interés. Encuentre el valor de u que minimiza la función

$$g(u) = \sum_{i=1}^{n} (x_i - u)^2.$$

51. Diga falso o verdadero.

    a) Sea $x_1, \ldots, x_n$ una colección de datos numéricos en donde cada uno de estos registros es igual a un mismo valor. La varianza de esta colección de datos es cero.

    b) Si la varianza de un conjunto de datos es cero, entonces todos los datos son idénticos.

    c) Si se le añaden ceros a un conjunto de datos, la varianza no cambia.

52. Sea $x_1, \ldots, x_n$ un conjunto de datos numéricos con media 4 y varianza 12. Encuentre la media y la varianza de los datos $y_1, \ldots, y_n$, en donde

    a) $y_i = 2x_i$.

    b) $y_i = -x_i + 1$.

    c) $y_i = x_i - 3$.

    d) $2x_i + 3y_i = 0$.

53. Sea $x_1, \ldots, x_n$ un conjunto de datos numéricos con media $\bar{x} \neq 0$. Defina $y_i = x_i/\bar{x}$ para $i = 1, 2, \ldots, n$. Compruebe que

$$s_y^2 = \frac{1}{\bar{x}^2} \cdot s_x^2.$$

54. La media y varianza de los salarios de un grupo de trabajadores es de \$550.00 y $(\$50)^2$, respectivamente. Si se aumentan los salarios en un 10 %, calcule la media y la varianza de los nuevos salarios.

Desviación estándar

55. Sea s_x la desviación estándar del conjunto de datos numéricos $x_1, \ldots, x_n$. Suponga que estos números se transforman en los datos $ax_1+c, \ldots, ax_n+c$, en donde a y c son dos constantes. Sean $y_1, \ldots, y_n$ los nuevos datos y sea s_y su desviación estándar. Compruebe que

$$s_y = |a| \cdot s_x.$$

56. Calcule la desviación estándar de los siguientes conjuntos de datos. La primera columna corresponde al conjunto de datos original. El resto de las columnas se obtiene según la operación indicada en el encabezado.

x	$\lvert x\rvert$	$x+2$	$3x-2$
-1	1	1	5
0	0	2	-2
0	0	2	-2
1	1	3	1

57. Sea $x_1, \ldots, x_n$ un conjunto de datos numéricos con media $\bar{x}$ y desviación estándar $s_x > 0$. Suponga que estos números se transforman en los datos $(x_1-\bar{x})/s_x, \ldots, (x_n-\bar{x})/s_x$. Sean $y_1, \ldots, y_n$ los nuevos datos. Compruebe que $\bar{y} = 0$ y $s_y^2 = 1$.

Desviación media

58. Calcule la desviación media de los siguientes conjuntos de datos. La primera columna corresponde al conjunto de datos original. El resto de las columnas se obtiene según la operación indicada en el encabezado.

x	$x+2$	$x-3$	$2x$	$x/2$	$-5x$
1	3	-2	2	1/2	-5
2	4	-1	4	1	-10
3	5	0	6	3/2	-15
3	5	0	6	3/2	-15
4	6	1	8	2	-20
5	7	2	10	5/2	-25

59. Suponga que los datos numéricos $x_1, \ldots, x_n$ se transforman en los datos $ax_1 + c, \ldots, ax_n + c$, en donde a y c son dos constantes. Compruebe que

$$\mathrm{dm}(ax + c) = |a|\,\mathrm{dm}(x).$$

Rango

60. Calcule el rango del siguiente conjunto de datos.

    a) $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$.

    b) $10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0$.

    c) $10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20$.

    d) $-10, -9, -8, -7, -6, -5, -4, -3, -2, -1, 0$.

61. Diga falso o verdadero.

a) El rango de un conjunto de datos puede ser cero.

b) El rango de un conjunto de datos puede ser negativo.

c) El rango de un conjunto de datos siempre es no negativo.

d) Dos conjunto de datos distintos pueden tener el mismo rango.

62. Suponga que los datos numéricos $x_1, \ldots, x_n$ se transforman en los datos $ax_1 + c, \ldots, ax_n + c$, en donde a y c son dos constantes. Compruebe que

$$\text{Rango}(ax + c) = |a| \cdot \text{Rango}(x).$$

Coeficiente de variación

63. Conteste las siguientes preguntas.

a) ¿Puede el coeficiente de variación ser negativo?

b) ¿Puede el coeficiente de variación ser cero?

c) ¿Qué se puede decir de un conjunto de datos cuyo coeficiente de variación es cero?

64. Sea $x_1, \ldots, x_n$ un conjunto de datos numéricos con media $\bar{x} \neq 0$. Defina $y_i = x_i/\bar{x}$ para $i = 1, 2, \ldots, n$. Compruebe que

$$\text{cv}(y) = \frac{1}{|\bar{x}|} \cdot s_x = \begin{cases} \text{cv}(x) & \text{si } \bar{x} > 0, \\ -\text{cv}(x) & \text{si } \bar{x} < 0. \end{cases}$$

Momentos

65. Sea m_k el k-ésimo momento central de un conjunto de datos $x_1, \ldots, x_n$. Compruebe que

a) $m_1 = 0$.

b) $m_2 = s^2$.

c) $m_2 = m_2' - (m_1')^2$.

66. Sea $m_k(x)$ el k-ésimo momento central de un conjunto de datos $x_1, \ldots, x_n$. Sea c una constante y considere la colección de datos trasladados $x_1 + c, \ldots, x_n + c$. Compruebe que

$$m_k(x + c) = m_k(x).$$

67. Sea $m_k(x)$ el k-ésimo momento central de un conjunto de datos $x_1, \ldots, x_n$. Sea a una constante y considere la colección de datos transformados $ax_1, \ldots, ax_n$. Compruebe que

$$m_k(ax) = a^k \cdot m_k(x).$$

Frecuencias

68. Suponga que se tiene una variable cualitativa ordinal con valores ordenados de menor a mayor A, B, C, D, E, F. Suponga además que una serie de observaciones de esta variable produce las frecuencias que aparecen en la siguiente tabla. Complete esta tabla calculando las frecuencias faltantes. Elabore además una gráfica de la frecuencia y otra de la frecuencia acumulada.

Valor	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
A	2			
B	8			
C	6			
D	4			
E	3			
F	5			

69. Suponga que se tiene una variable cualitativa ordinal con valores ordenados de menor a mayor A, B, C, D. Suponga además que una serie

de 20 observaciones de esta variable produce las frecuencias relativas que aparecen en la siguiente tabla. Complete esta tabla calculando las frecuencias faltantes. Elabore además una gráfica de la frecuencia y otra de la frecuencia acumulada.

Valor	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
A			0.1	
B			0.4	
C				
D			0.2	

Cuantiles

70. Explique la diferencia entre

    a) un cuantil y un cuartil.

    b) un cuantil y un percentil.

71. Calcule el cuantil al 25 %, al 50 % y al 75 % del siguiente conjunto de datos.

    a) $-1, 0, 1$

    b) $1, 2, 2, 2, 2, 3.$

    c) $-1, 5, 0, 3, 3, 2, 1, 0.$

    d) $2.5, 3.2, 2.7, 2.0, 3.4, 3.7, 5.2, 2.0, 2.5, 4.0, 2.3.$

    e) $20, 32, 25, 27, 30, 28, 31, 20, 30.$

72. Indique en cada uno de los dos conjuntos de datos que aparecen en la Figure 2.22 los cuantiles al 20 %, 40 %, 60 % y 80 %.

Figura 2.22

Coeficiente de asimetría (*Skewness*)

73. Sea $\text{sk}(x)$ el coeficiente de asimetría del conjunto de datos numéricos $x_1, \ldots, x_n$. Sea c una constante y considere los datos trasladados $x_1 + c, \ldots, x_n + c$. Compruebe que el coeficiente de asimetría no cambia, es decir,

$$\text{sk}(x + c) = \text{sk}(x).$$

74. Sea $\text{sk}(x)$ el coeficiente de asimetría del conjunto de datos numéricos $x_1, \ldots, x_n$. Sea a una constante distinta de cero y considere los datos transformados $ax_1, \ldots, ax_n$. Compruebe que

$$\text{sk}(ax) = \begin{cases} \text{sk}(x) & \text{si } a > 0, \\ -\text{sk}(x) & \text{si } a < 0. \end{cases}$$

Curtosis

75. Calcule la curtosis del siguiente conjunto de datos.

    a) $-1, 0, 1$.

    b) $-2, 0, 2$.

    c) $0, 1, 2$.

76. Sea $k(x)$ la curtosis de un conjunto de datos $x_1, \ldots, x_n$. Sea c una constante y considere la colección de datos trasladados $x_1+c, \ldots, x_n+c$ con curtosis $k(x + c)$. Compruebe que

$$k(x + c) = k(x).$$

77. Sea $k(x)$ la curtosis de un conjunto de datos $x_1, \ldots, x_n$. Sea a una constante distinta de cero y considere la colección de datos transformados $ax_1, \ldots, ax_n$ con curtosis $k(ax)$. Compruebe que

$$k(ax) = k(x).$$

Datos agrupados

78. Calcule la media, la moda y la mediana del siguiente conjunto de datos agrupados. Utilice alguna de las dos perspectivas de selección de la marca de clase.

Clase	Frecuencia
$C_1 = \{0\}$	$f_1 = 3$
$C_2 = \{1, 2\}$	$f_2 = 5$
$C_3 = \{3, 4\}$	$f_3 = 2$
$C_4 = \{5, 6\}$	$f_4 = 4$

Capítulo 3

Descripciones gráficas

Hasta un cierto nivel de precisión, la información representada mediante gráficas es más fácil de entender para la mente humana. Por ejemplo, las gráficas son de gran ayuda para comprender rápidamente la comparación en magnitud entre dos o mas cantidades, o bien en un gráfica se puede observar con claridad un posible comportamiento creciente (o decreciente) de una cierta cantidad, se puede tener también una idea de la velocidad del crecimiento (o del decrecimiento) de una variable, e incluso de manera intuitiva la mente humana puede hacer predicciones del comportamiento de la cantidad en estudio. Todo esto posible y es motivado por la representación gráfica de la información. Nuestra mente, además, lleva a cabo estas acciones en fracciones de segundo y sin aparente mayor determinación nuestra por hacerlo. Parece ser que estamos propensos a ello.

Teniendo esta justificación en mente, en este capítulo se explicarán algunas formas de representar gráficamente la información de un conjunto de datos. Las gráficas que construiremos no son todas las existentes, en realidad muchas otras formas imaginativas de gráficas pueden proponerse. Estas gráficas tienen el objetivo de transmitir información de manera rápida, sencilla, resumida y de fácil e inmediata comprensión.

La elaboración de gráficas se facilita mucho mediante el uso de programas de cómputo para el tratamiento de datos. Estos programas pueden producir gráficas muy atractivas y con múltiples opciones para su configuración. Sin

embargo, es recomendable buscar siempre claridad y facilidad de lectura en el diseño de una gráfica. Los ejemplos gráficos que presentaremos en este capítulo son modestos y tienen la finalidad de mostrar de manera básica algunos tipos de gráficas. Dependiendo de cada programa de cómputo utilizado, pueden existir muchas variaciones y mejoramientos estéticos en la presentación del tipo de gráficas que estudiaremos. Para el paquete `R`, por ejemplo, se puede consultar el libro [6].

Gráfica de barras

Esta es una gráfica simple que consiste de varias barras que representan las categorías (o agrupamiento de valores) de una variable. En el eje horizontal se colocan las categorías, se dibuja una barra para cada categoría y la altura de la barra es la frecuencia o número de veces que se observa la categoría. El ancho de la barra no es relevante y puede no ser homogéneo para todas las categorías, en caso de que se decida agrupar algunas de ellas.

Para este tipo de gráficas, la variable puede ser cualitativa o cuantitativa, y las categorías o agrupamiento de valores pueden ser nominales u ordinales. En el caso de variable nominales, las clases o categorías se pueden colocar en cualquier orden.

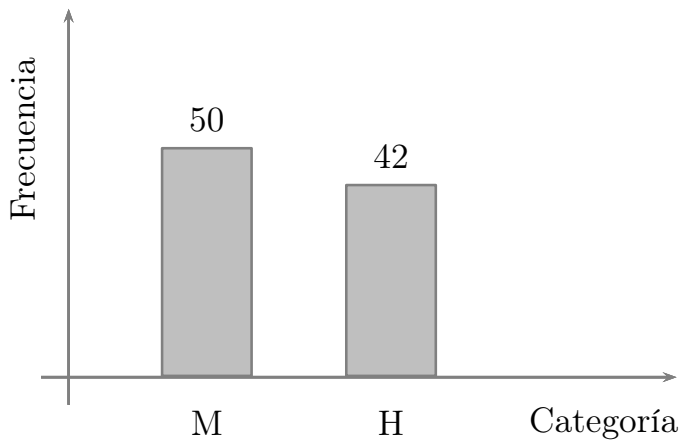


Figura 3.1: Ejemplo de una gráfica de barras de frecuencias para una variable categórica.

Veamos un ejemplo. Supongamos que tenemos un conjunto de 92 personas, de las cuales 50 son mujeres (M) y 42 son hombres (H). Esta información puede representarse mediante una gráfica de barras como en la Figura 3.1.

En este ejemplo las categorías M y H no tienen orden y pueden ser colocadas en la gráfica en cualquier orden conveniente. Además, las barras pueden separarse un poco, como en el ejemplo, o colocarse una contigua a la otra. Como información adicional en la gráfica es conveniente colocar los valores numéricos de las frecuencias en la parte superior de cada barra.

Las gráficas de barra ayudan a visualizar los valores de una variable que ocurren con mayor o menor frecuencia y a comparar cualitativamente estas frecuencias.

Como un segundo ejemplo considere la tabla con la frecuencia de los grupos sanguíneos de 80 personas que se muestra en la Tabla 3.1.

Grupo Sanguíneo	Frecuencia	Porcentaje
A	12	15 %
B	28	35 %
AB	16	20 %
O	24	30 %

Tabla 3.1: Ejemplo de frecuencias para una variable categórica.

Las frecuencias que aparecen en esta tabla se pueden representar mediante el histograma que aparece en la Figura 3.2.

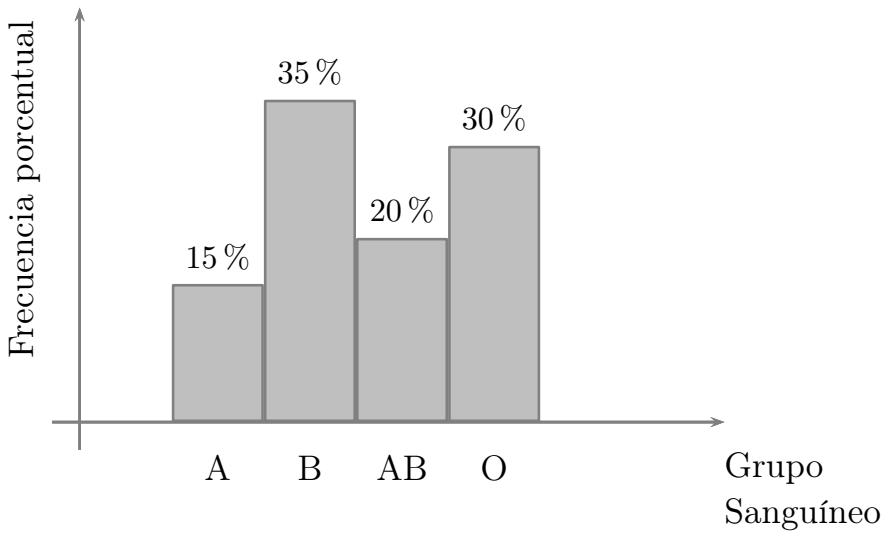


Figura 3.2: Ejemplo de gráfica de barras de frecuencias para una variable categórica.

En este caso se han colocado las barras de manera contigua y se han indicado las frecuencias porcentuales en la parte superior de cada barra. Observe que no se pueden conocer las frecuencias absolutas a menos que se indique que las observaciones se obtuvieron de 80 personas.

En el paquete R se puede obtener una gráfica de barras similar a la que hemos presentado. Se usa la siguiente serie de instrucciones y el resultado se muestra en la Figura 3.3.

```
> f <- c(15,35,20,30)
> gs <- c("A","B","AB","O")
> barplot(f,names.arg=gs,
xlab="Grupo sanguineo",ylab="Frecuencia")
```

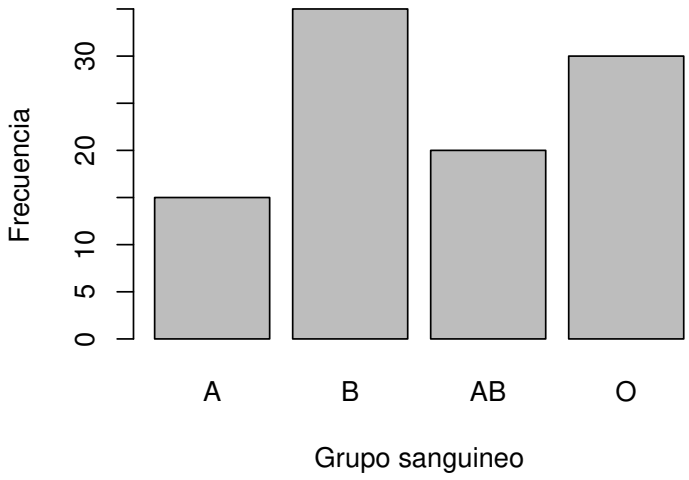


Figura 3.3: Ejemplo de gráfica de barras producida en el paquete R.

Las gráficas de barras también pueden presentarse en forma horizontal. Por ejemplo, según la información de *The World Factbook* para mediados del año 2017, los 10 principales países productores de petróleo fueron los que se muestran en la Figura 3.4. Es evidente que este tipo de representaciones gráficas ayudan a la rápida comprensión de la información numérica.

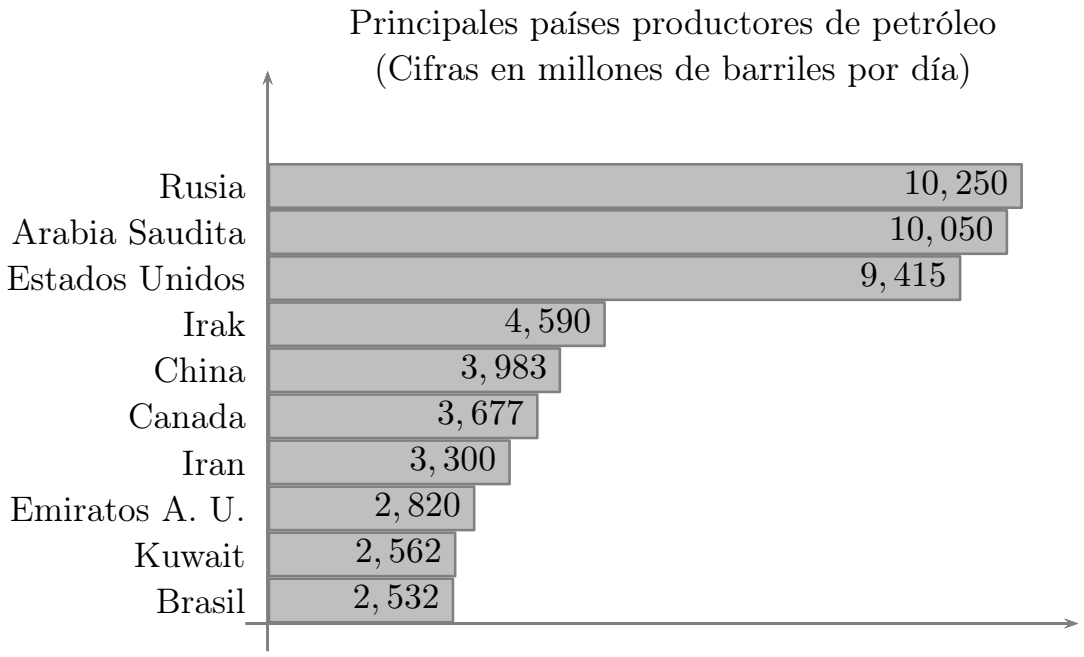


Figura 3.4: Ejemplo de una gráfica de barras horizontal.

Histograma

Un histograma es una gráfica muy similar a una gráfica de barras. Adquiere este nombre cuando existe un orden entre los valores de la variable a graficar. Salvo esta condición, los datos puede ser cualitativos o cuantitativos. Nuevamente, para cada valor, categoría o clase de la variable, se asocia una barra cuya altura es la frecuencia con la que se observa la categoría. Como las categorías tienen un orden, se representan regularmente en el eje horizontal de menor a mayor. Como en las gráficas de barras y para mayor información, en la parte superior de cada barra se puede colocar la frecuencia absoluta, la frecuencia relativa o la frecuencia porcentual.

A menudo se tiene una gran cantidad de datos numéricos y para elaborar un histograma con esta información se definen agrupaciones de valores, en este caso intervalos, y se calcula el número de datos que quedan en cada intervalo. A partir de estas gráficas se pueden sugerir modelos teóricos de

probabilidad para la variable en estudio. Refinando o engrosando los intervalos de valores pueden obtenerse histogramas más claros y sugerentes. Por ejemplo, en la Figura 3.5 se muestra un histograma que claramente se asemeja a la conocida curva en forma de campana y sugiere, por lo tanto, que la variable en estudio puede adoptar el modelo teórico de esa curva. Esta es la famosa distribución normal o distribución gaussiana.

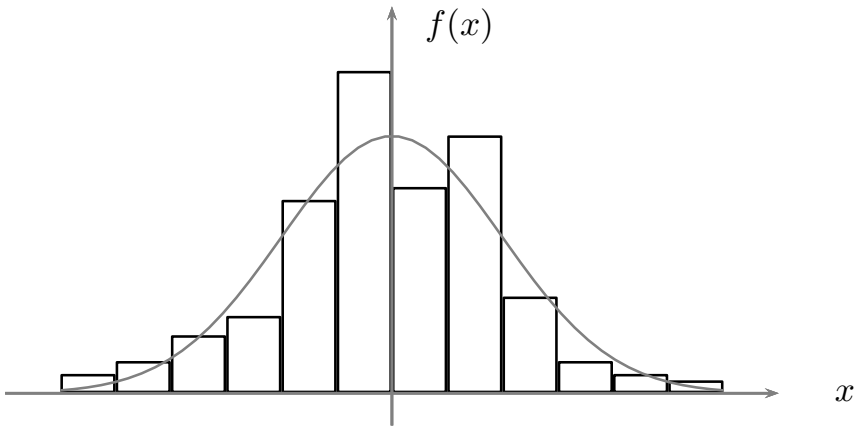


Figura 3.5: Histograma semejante a una curva en forma de campana.

Se pueden elaborar también histogramas de frecuencias de ocurrencias de un cierto evento a lo largo del tiempo. En este caso en el eje horizontal se colocan los intervalos de tiempo y en el eje vertical las frecuencias observadas.

Polígono de frecuencias

Para construir un polígono de frecuencias se marcan los puntos medios en la parte superior de las barras de un histograma y se unen estos puntos a través de líneas rectas. A la gráfica resultante se le llama polígono de frecuencias. En la Figura 3.6 se muestra un ejemplo de este tipo de gráficas.

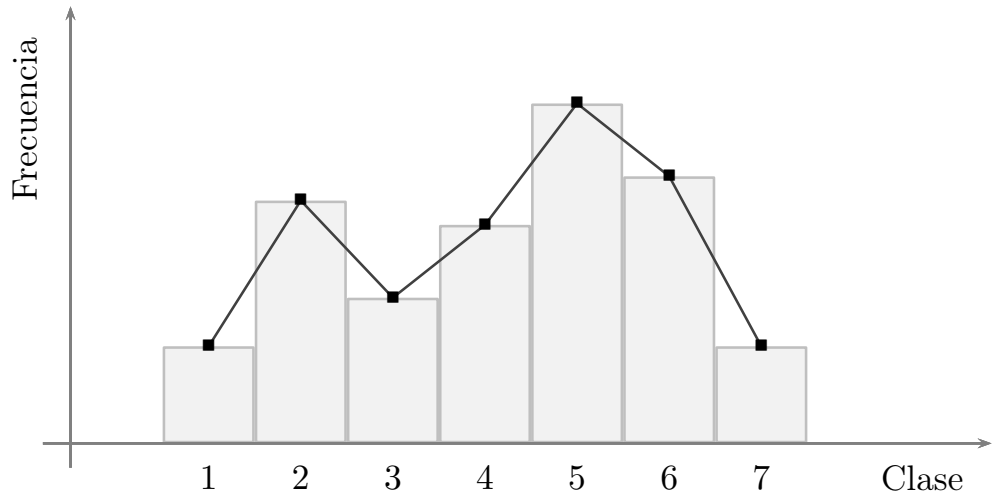


Figura 3.6: Ejemplo de un polígono de frecuencias.

La información presentada mediante un polígono de frecuencias es equivalente a la información de un histograma, sin embargo, dado que se trata de líneas rectas, las tendencias de crecimiento y decrecimiento son más evidentes. Esta es una de las utilidades de este tipo de gráficas.

Polígono de frecuencias acumuladas

Esta es una gráfica equivalente al histograma de frecuencias acumuladas. Para su construcción se marcan los puntos medios en la parte superior de las barras de un histograma de frecuencias acumuladas. Nuevamente se unen estos puntos a través de líneas rectas. A la gráfica resultante se le llama polígono de frecuencias acumuladas. En la Figura 3.7 se muestra un ejemplo de este tipo de gráficas.

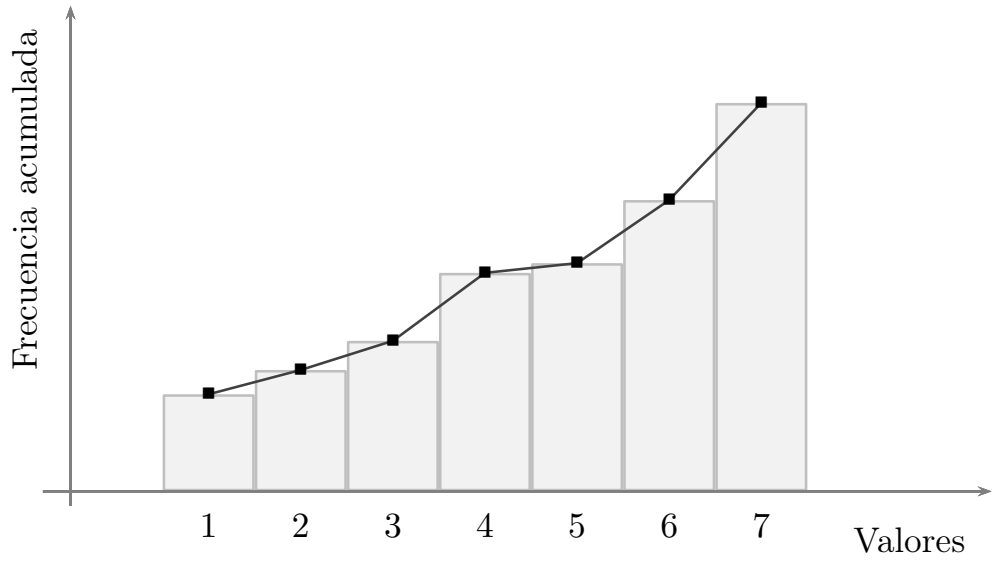


Figura 3.7: Ejemplo de un polígono de frecuencias acumuladas.

Es evidente que el comportamiento creciente de las frecuencias acumuladas es más claramente identificado en este tipo de gráficas. Para mayor información en la gráfica y, si resulta conveniente, se pueden colocar los valores numéricos de las frecuencias acumuladas arriba del punto marcado en cada barra.

Ojiva

Una ojiva es una curva suave que se traza sobre los puntos de un polígono de frecuencias acumuladas. Se aplica para clases o agrupamientos de valores ordinales y la curva resultante es más fácil de dibujar cuando el número de clases es grande. La ojiva es una idealización del comportamiento creciente del polígono de frecuencias acumuladas. En la Figura 3.8 se muestra una de estas gráficas.

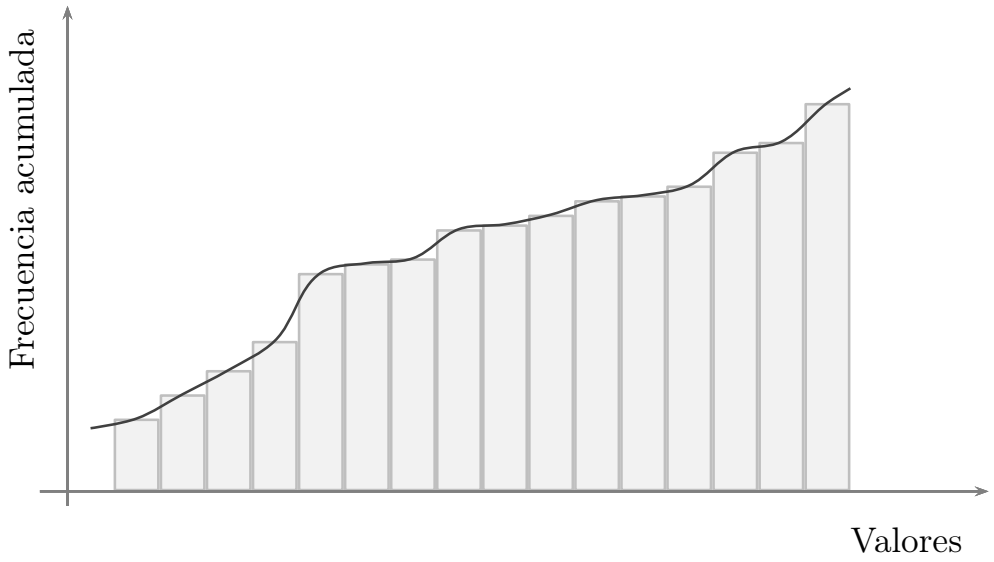


Figura 3.8: Ejemplo de una ojiva.

Gráfica de pastel

Para variables cualitativas o bien para variables cuantitativas agrupadas, se pueden elaborar gráficas de pastel, también llamadas *pie charts*. Estas son gráficas circulares divididas en sectores que permiten comparar visualmente las frecuencias porcentuales de los valores observados de una variable.

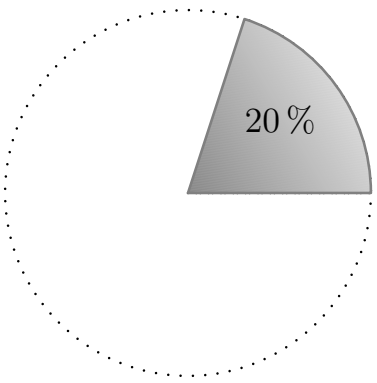


Figura 3.9: Ejemplo de un sector.

La frecuencia de una categoría o grupo de valores se representa mediante un sector de una circunferencia cuyo ángulo se determina de la siguiente forma: una frecuencia relativa, por ejemplo, de 0.2 (véase la Figura 3.9) se asocia con un sector con un ángulo de

$$(0.2) \times (360^\circ) = 72^\circ.$$

Veamos un ejemplo. Consideremos nuevamente el grupo de 80 personas en donde se ha registrado su grupo sanguíneo. Reproducimos nuevamente esta información en la Tabla 3.2 y con esta información elaboraremos una gráfica de pastel y por ello se ha añadido en la tabla el ángulo del sector. Con esta información se puede construir la gráfica de pastel que aparece en la Figura 3.10.

Grupo Sanguíneo	Frecuencia	Porcentaje	Ángulo del sector
A	12	15 %	54°
B	28	35 %	126°
AB	16	20 %	72°
O	24	30 %	108°

Tabla 3.2: Ejemplo del cálculo de ángulos de sectores.

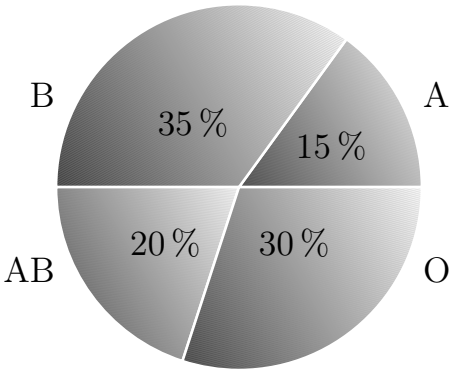


Figura 3.10: Ejemplo de una gráfica de pastel.

Consideremos otro ejemplo. Supongamos que tenemos la información de un equipo de fútbol que en la temporada ha jugado 16 partidos, de los cuales ha perdido 2 (12.5 %), empatado 4 (25 %), y ganado 10 (62.5 %). Se puede representar esta información mediante la gráfica que se muestra en la Figura 3.11.

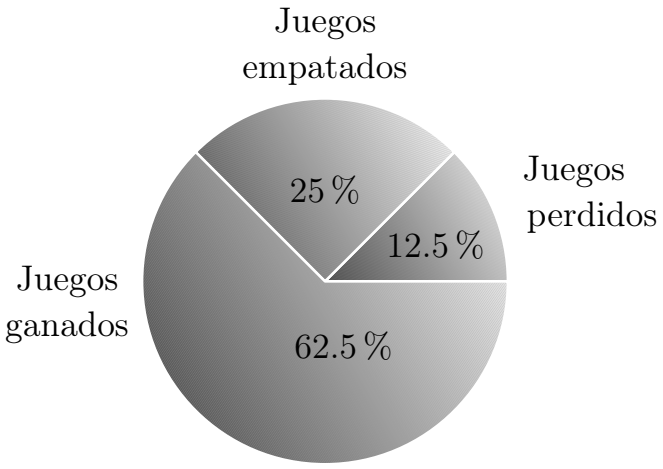


Figura 3.11: Ejemplo de una gráfica de pastel.

Visualmente las gráficas de pastel son vistosas y logran muy bien el propó-

sito de resumir la información en una gráfica. Se pueden dibujar gráficas de pastel en tercera dimensión y usar colores para hacerlas aún más atractivas. A continuación mostramos los comandos básicos con los cuales se pueden obtener gráficas de pastel en `R`. Reproduciremos los dos ejemplos mencionados. Las gráficas se muestran en la Figura 3.12. Para el primer ejemplo tenemos las siguientes instrucciones.

R

```
> sectores <- c(15,35,20,30)
> categorias <- c("A","B","AB","O")
> pie(sectores, labels = categorias)
```

Para el segundo ejemplo, se puede generar una gráfica básica en tercera dimensión de la siguiente forma, suponiendo instalada la biblioteca `plotrix`.

R

```
> library(plotrix)
> sectores <- c(2,4,10)
> categorias <- c("JP", "JE", "JG")
> pie3D(sectores, labels=categorias, explode=0.1)
```

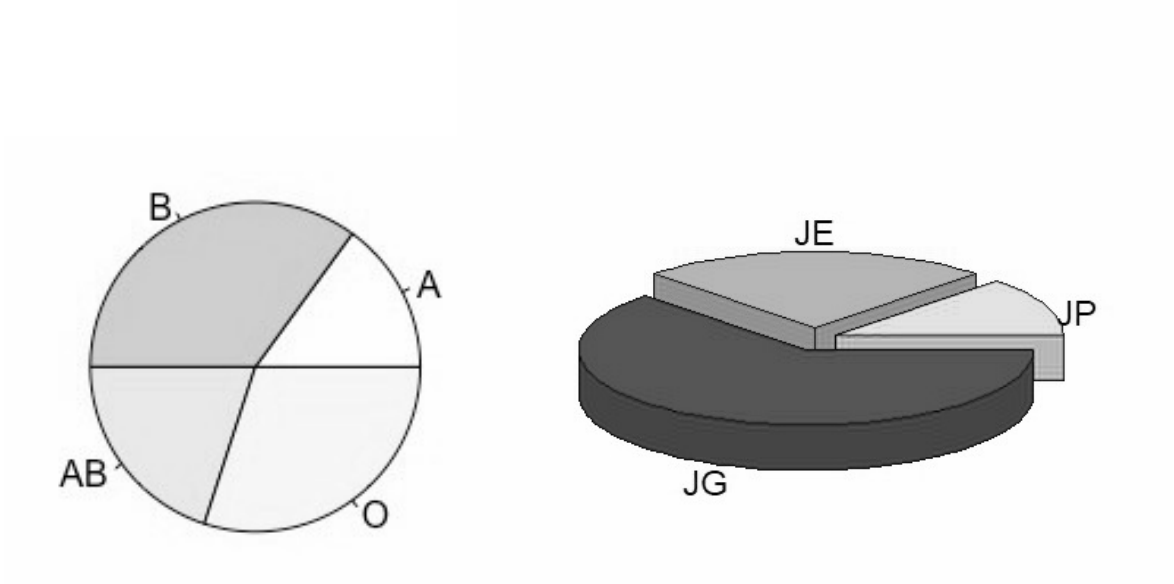


Figura 3.12: Ejemplos de gráficas de pastel producidas en el paquete `R`.

Recordemos que nuestra intención es proveer ejemplos introductorios de comandos básicos en R para producir gráficas simples. Si se desea mejorar las gráficas presentadas, se deberá consultarse algún manual de R para conocer las distintas opciones disponibles.

Gráfica de tallo y hojas

Esta es otra forma gráfica de representar un conjunto de datos numéricos. Su aspecto es muy similar al de un histograma dibujado horizontalmente. Daremos varios ejemplos para ilustrar la construcción de este tipo de gráficas.

Consideremos el siguiente conjunto de datos

126	102	84	100	67	89
73	124	113	91	92	96
112	70	82	95	121	126
72	84	87	92	107	100

A continuación se separa el dígito menos significativo del resto de los dígitos mediante una línea vertical, por ejemplo, el primer valor 126 se separa en $12\,|\,6$. Se puede entonces conformar un diagrama como se muestra en la Figura 3.13, en donde se han ordenado los dígitos separados de menor a mayor, incluyendo repeticiones.

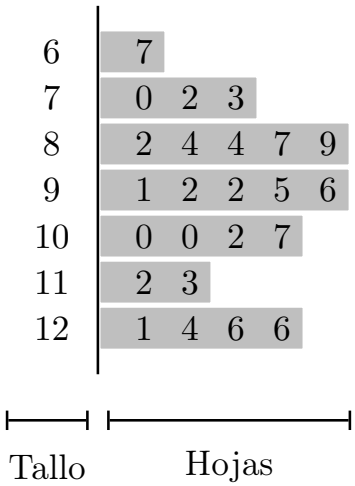


Figura 3.13: Ejemplo de gráfica de tallo y hojas.

Este es un diagrama de tallo y hojas. A los dígitos separados y que aparecen en la parte derecha del diagrama se les llama hojas y a la parte izquierda se le llama tallo. Si este diagrama se rota 90 grados en el sentido contrario al movimiento de las manecillas del reloj, se obtiene un diagrama similar al de un histograma. En general, los datos deben ser cercanos unos a otros para que los diagramas de tallo y hojas resultantes tengan una forma compacta.

En el paquete `R` se pueden obtener gráficas de tallo y hojas usando el comando `stem()`. La palabra en inglés *stem* se traduce como *tallo*. A continuación repetimos el ejemplo mostrado pero ahora en `R`.

R

```
> x<-c(126,102,84,100,...,72,84,87,92,107,100)
> stem(x,scale=2)
```

El resultado que arroja `R` es el siguiente:

```
The decimal point is 1 digit(s) to the right of the |
```

```
 6 | 7
 7 | 023
 8 | 24479
 9 | 12256
10 | 0027
11 | 23
12 | 1466
```

Variantes

- Si algún tallo tiene demasiadas hojas se pueden separar las hojas en varias partes, por ejemplo, véase la Figura 3.14, en donde existen muchos datos entre los valores 70 y 80, y se han separado en dos grupos.

6	8 9
7	2 3 3
7	5 6 9
8	0 1

Figura 3.14: Ejemplo de separación de un tallo (valor 7) en dos partes.

- Si resulta conveniente, los datos con muchos dígitos se pueden recortar, por ejemplo, para el conjunto de números

2104	1757	1562	1756	1730
1992	1683	2133	2013	1684
1710	1881	1961	1672	1855

el primer dato 2104 se puede recortar a 210 y la separación es $21 \,|\, 0$. Se elabora entonces el diagrama de la Figura 3.15 indicando que la unidad de la hoja es 10. En este caso se pierde precisión de los datos originales pero se gana simplicidad en la presentación.

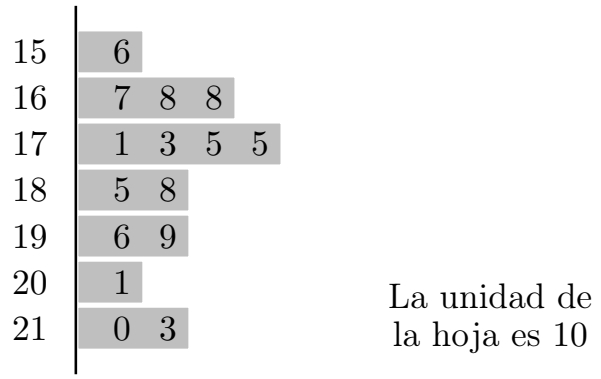


Figura 3.15: Ejemplo en donde se han recortado los datos a los tres dígitos más significativos.

La unidad de la hoja indica el número por el que debe multiplicarse el dato graficado para obtener una aproximación del dato original. Por ejemplo, el primer dato graficado $15\,|\,6$ en la Figura 3.15 corresponde a un valor aproximado de 1560. La unidad de las hojas puede ser 100, 10, 1, 0.1, 0.01, etcétera. Por ejemplo, si la unidad de la hoja es 0.1, el dato graficado $15\,|\,6$ corresponde al valor 15.6 .

Diagrama de caja y brazos

Esta es una forma gráfica de representar de manera resumida un conjunto de datos numéricos $x_1, \ldots, x_n$. Esta representación está compuesta por una caja y por dos marcas en dos de sus extremos opuestos que asemejan brazos como se muestra en la Figura 3.16 . A este tipo de gráficas se les conoce también como diagramas de caja y bigotes, y por los términos en inglés *boxplots* o *whiskers*. Para dibujar estos diagramas se necesita determinar el centro de la caja, su altura y los tamaños de los brazos superior e inferior. Explicaremos dos maneras en las que se pueden determinar estos parámetros.

Para el ejemplo mostrado en la Figura 3.16, el centro de la caja es la media $\bar{x}$, se extiende la caja una desviación estándar s hacia arriba y otra desviación estándar s hacia abajo. La caja tiene, por lo tanto, una altura de $2s$

unidades. La marca del brazo superior es igual al máximo valor observado, esto es, $x_{(n)}$. La marca del brazo inferior es el mínimo valor observado, es decir, $x_{(1)}$. En general, las longitudes de los brazos son distintas.

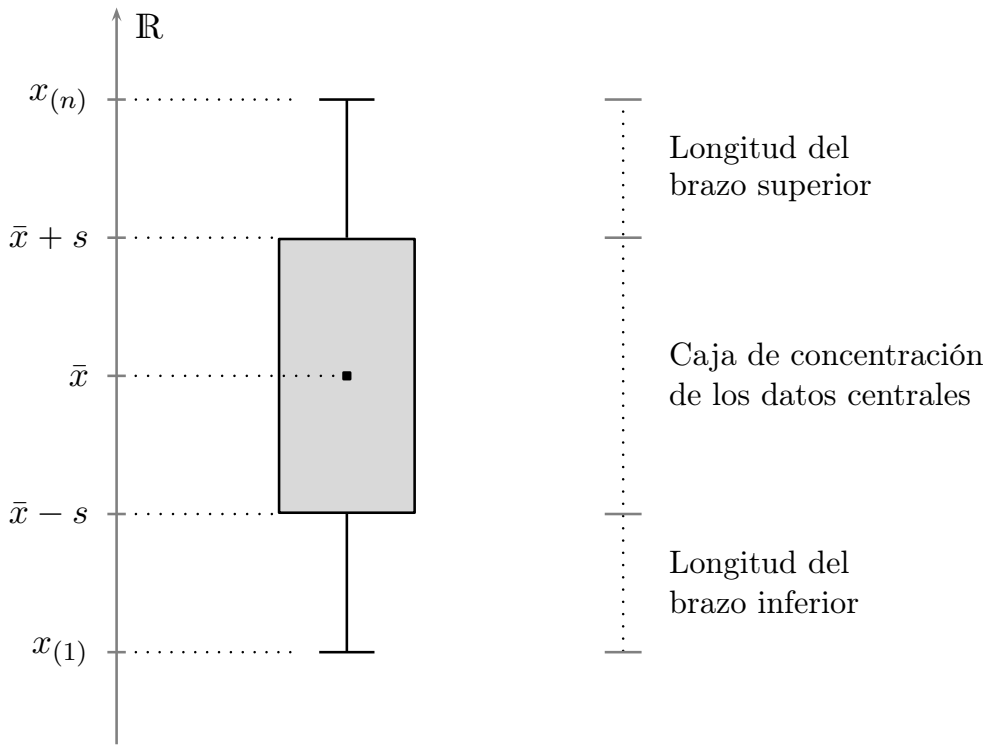


Figura 3.16: Una primera forma de definir una gráfica de caja y brazos.

De esta manera, un diagrama de caja y brazos, construido de la forma indicada, es una forma de representar 4 descripciones numéricas de un conjunto de datos en un solo diagrama: el dato menor $x_{(1)}$, la media $\bar{x}$, la desviación estándar s, y el dato mayor $x_{(n)}$. Se pueden colocar dos o más de estos diagramas, uno junto al otro, a fin de comparar visualmente estas características en distintos conjunto de datos.

Otra manera de construir un diagrama de caja y brazos es a través de los cuantiles. La altura de la caja parte del primer cuartil $Q_{0.25}$ y se extiende hasta el tercer cuartil $Q_{0.75}$. Observe que el segundo cuartil $Q_{0.5}$, es decir, la mediana, se encuentra dentro de la caja pero no necesariamente en el centro.

La altura de la caja es entonces el así llamado **rango intercuartil**

$$\text{RIC} = Q_{0.75} - Q_{0.25}.$$

Véase la Figura 3.17. El rango intercuartil mide la longitud del intervalo más pequeño que contiene el 50 % de los datos centrales alrededor de la mediana. Por su nombre en inglés, el rango intercuartil también puede denotarse por las letras IQR, *interquartile range*. Las longitudes de los brazos se puede establecer como 1.5 veces el rango intercuartil RIC, y en este caso, los brazos tienen idéntica longitud.

A los valores que se encuentren abajo de la marca del brazo inferior o arriba de la marca del brazo superior se les llama valores atípicos. A los valores que se encuentren arriba de $Q_{0.75} + 3\text{RIC}$ o abajo de $Q_{0.25} - 3\text{RIC}$ se les llama extremadamente atípicos (*outliers*).

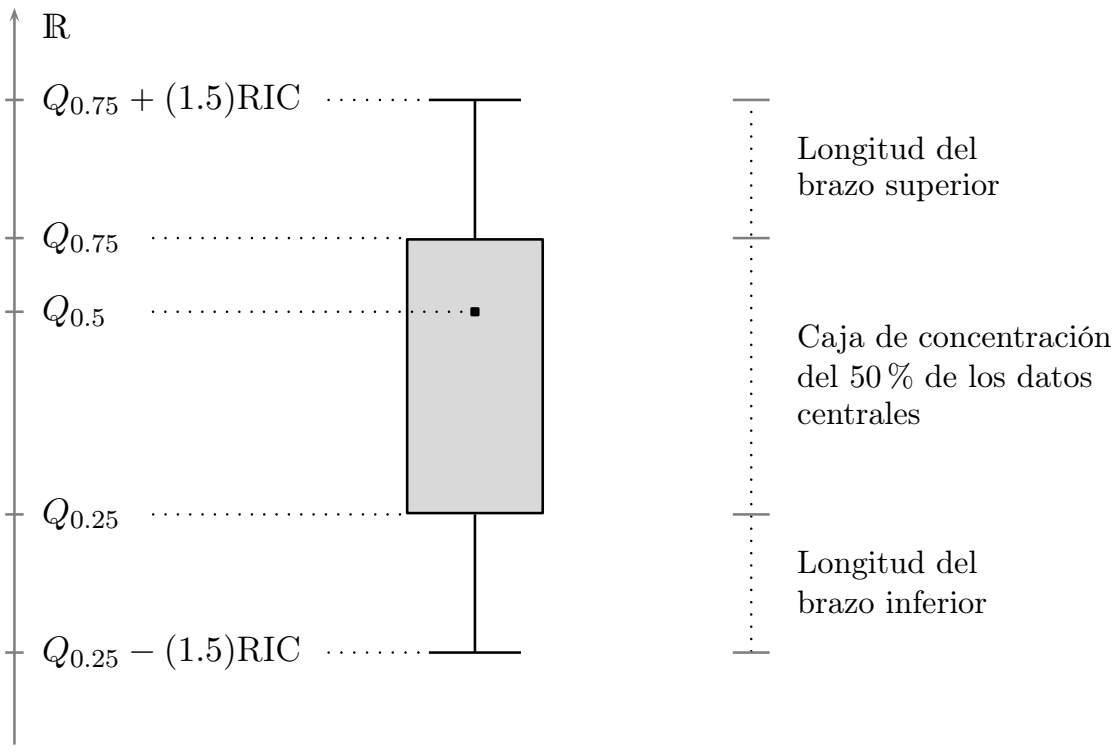


Figura 3.17: Una segunda forma de definir una gráfica de caja y brazos.

Las marcas de los brazos inferior y superior pueden ser los cuantiles al 10 % y 90 %, respectivamente, o bien los cuantiles al 5 % y 95 %.

El comando para producir gráficas de caja y brazos en `R` es `boxplot()`. La forma de definir la caja es a través de los cuartiles pero la manera de calcular la longitud de los brazos es un poco distinta a la explicada arriba. A continuación mostramos un ejemplo sencillo con algunos pocos datos.

R

```
> x <- c(-5,1,2,3,4,5,6,7,8,9,10,15)
> boxplot(x)
```

El resultado se muestra en la Figura 3.18. Es importante mencionar que, de manera estándar, `R` usa las siguientes cantidades para conformar la gráfica de caja y brazos: la caja tiene longitud el rango intercuartil RIC $= Q_{0.75} - Q_{0.25}$, la marca dentro de la caja es la mediana $Q_{0.50}$, y los brazos se determinan de la siguiente forma

$$\begin{aligned} \text{Brazo superior} &= \text{mín}\,\{x_{(n)}, Q_{0.75} + 1.5 \cdot RIC\}, \\ \text{Brazo inferior} &= \text{máx}\,\{x_{(1)}, Q_{0.25} - 1.5 \cdot RIC\}. \end{aligned}$$

Para el ejemplo mostrado en el recuadro, puede comprobarse que $Q_{0.25} = 2.5$, $Q_{0.50} = 5.5$ y $Q_{0.75} = 8.5$, de modo que RIC $= 6$. Así, el brazo superior es

$$\text{mín}\,\{15, 8.5 + 1.5 \cdot 6\} = \text{mín}\,\{15, 17.5\} = 15.$$

El brazo inferior es

$$\text{máx}\,\{-5, 2.5 - 1.5 \cdot 6\} = \text{máx}\,\{-5, -6.5\} = -5.$$

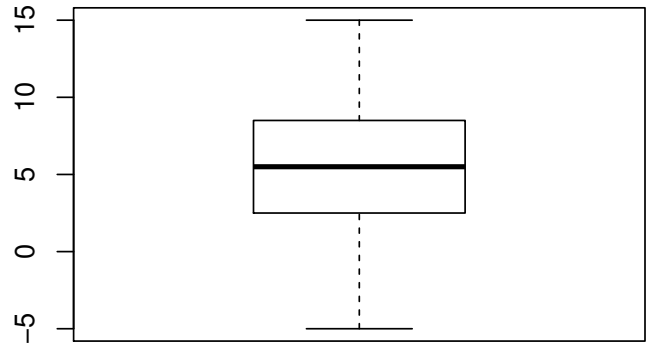


Figura 3.18: Ejemplo de gráfica de caja y brazos producida en el paquete R.

Función de distribución empírica

Sean $x_1, \ldots, x_n$ una sucesión de observaciones de una variable numérica de interés. La función de distribución empírica es otra manera gráfica de representar estas observaciones. Se denota por $F(x)$ (aquí es importante que sea la letra F sea en mayúscula) y se define de la forma siguiente.

$$F(x) = \frac{\#\{x_i \,:\, x_i \leqslant x\}}{n}$$

Recordemos que el símbolo # indica cardinalidad o número de elementos en el conjunto indicado. En esta definición el conjunto indicado consta de todas las observaciones x_i que se encuentran a la izquierda del valor x, incluyendo este valor. Esta es la razón por la que a la función $F(x)$ se le conoce también como la función de distribución empírica acumulada. De esta manera, para cada número real x se debe contar el número de observaciones que son menores o iguales a x (u observaciones acumuladas hasta x) y dividir entre el total de observaciones n.

Las gráficas de estas funciones tienen el aspecto de una escalera, presentando un escalón en cada observación x_i y en donde el tamaño del escalón es la frecuencia relativa del dato x_i. De esta manera en la función de distribu-

ción empírica está representada toda la información de la colección de datos numéricos. Veamos un ejemplo.

Ejemplo 3.1 Supongamos que tenemos las siguientes cuatro observaciones numéricas de una cierta variable de interés:

$$2, 1, 2, 3.$$

Hemos escogido estos pocos valores pues con ellos es suficiente para ilustrar la construcción de la función de distribución empírica. Siguiendo la definición anterior, puede comprobarse que esta función es

$$F(x) = \begin{cases} 0 & \text{si } x < 1, \\ 1/4 & \text{si } 1 \leqslant x < 2, \\ 3/4 & \text{si } 2 \leqslant x < 3, \\ 1 & \text{si } x \geqslant 3. \end{cases}$$

La forma de obtener esta función es la siguiente: notamos primero que si tomamos cualquier valor $x < 1$, no hay ninguna observación x_i tal que $x_i \leqslant x$. Por lo tanto, la función vale 0 para esos valores de x. Para cualquier x entre el valor 1 (inclusive) y el valor 2, hay siempre una observación antes que x, esa es la observación 1 y por lo tanto la función vale $1/4$ para esos valores de x. Si ahora tomamos x entre 2 (inclusive) y 3, hay siempre tres observaciones antes que x, éstas son $1, 2, 2$, la función entonces vale $3/4$ para esos valores de x. Finalmente, si tomamos x mayor o igual a 3 tenemos que todas las observaciones x_i son tales que $x_i \leqslant x$ y por lo tanto la función de distribución empírica vale uno para esos valores de x. La gráfica de esta función se muestra en la Figura 3.19.

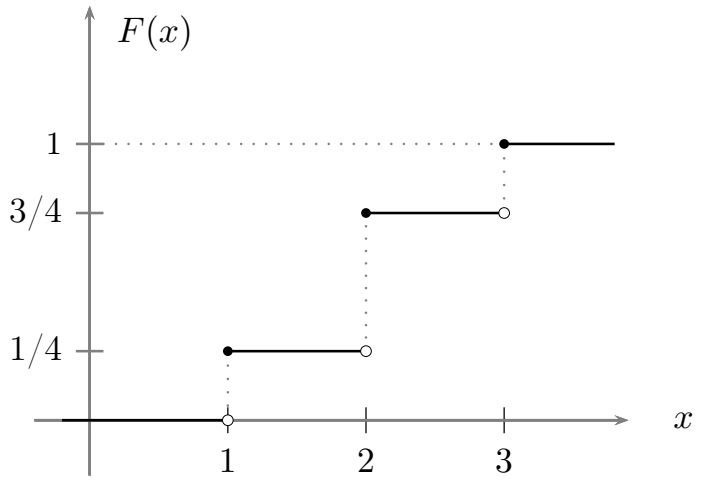


Figura 3.19: Ejemplo de una función de distribución empírica.

Observemos que como el dato 2 aparece dos veces, el escalón allí es de magnitud 2/4. Si todos los datos hubieran sido distintos, tendríamos una función de distribución empírica con cuatro escalones de magnitud 1/4 cada uno de ellos. •

Así, la función de distribución empírica inicia en el valor cero y se va incrementando mediante saltos hasta llegar al valor uno. En general, mientras mayor sea el número de datos observados, la función de distribución empírica toma un aspecto cada vez más parecido a una curva continua creciente. En alguna situación real, en donde se tenga una gran cantidad de datos, es necesario el uso de una computadora para graficar esta función. Por ejemplo, en el paquete `R` puede usarse el comando `ecdf(x)`, que proviene del término *empirical cumulative distribution function*. Mostramos esto a continuación para el ejemplo mostrado.

R

```
> x <- c(2,1,2,3)
> plot(ecdf(x))
```

Los resultados se muestran en la Figura 3.20. Esta gráfica es similar a la presentada en la Figura 3.19. Es claro que la primera gráfica es más precisa y limpia.

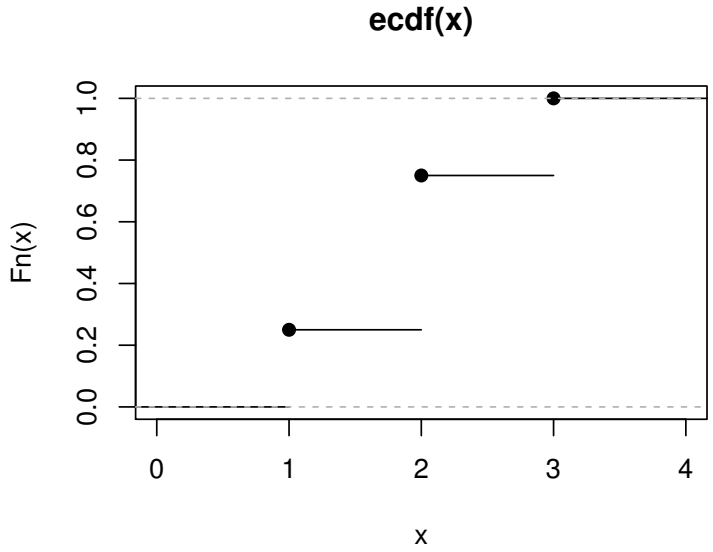


Figura 3.20: Función de distribución empírica producida en el paquete R.

Puede verificarse que toda función de distribución $F(x)$ satisface las siguientes propiedades:

- $F(x) = 0$ para cualquier $x < x_{(1)}$.
- $F(x) = 1$ para cualquier $x \geqslant x_{(n)}$.
- $F(x)$ es creciente, esto significa que si $x \leqslant y$ entonces $F(x) \leqslant F(y)$.
- $F(x)$ es continua por la derecha.

La función de distribución empírica es importante dentro de la probabilidad y la estadística en general, puesto que, desde el punto de vista teórico, en ella está contenida toda la información obtenida de las observaciones de la variable de interés.

Como un ejemplo de aplicación de la función de distribución empírica, a continuación vamos a explicar otra forma de calcular los cuantiles de una serie de datos a partir de su función de distribución empírica.

Ejemplo 3.2 Supongamos que tenemos las siguientes cuatro observaciones numéricas de una cierta variable:

$$1, 0, 2, 0.$$

En un orden distinto pero estos son los mismos valores considerados en el Ejemplo 2.2 para el cálculo de cuantiles que aparece en la página 66. Siguiendo la definición, la correspondiente función de distribución empírica tiene la siguiente expresión

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ 2/4 & \text{si } 0 \leqslant x < 1, \\ 3/4 & \text{si } 1 \leqslant x < 2, \\ 1 & \text{si } x \geqslant 2. \end{cases}$$

La gráfica de esta función se muestra en la Figura 3.21.

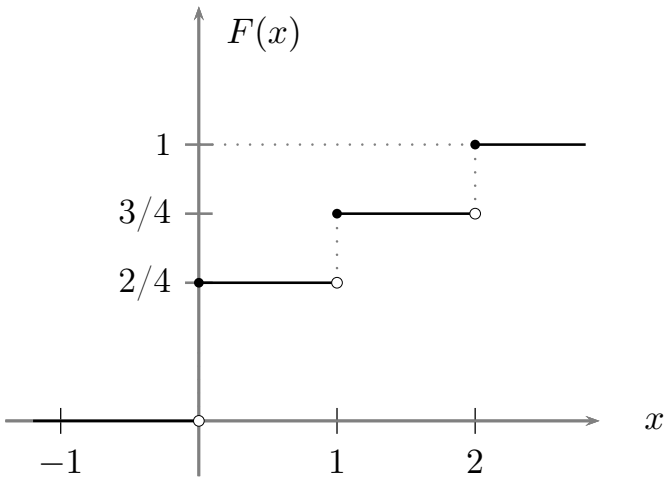


Figura 3.21: Ejemplo de una función de distribución empírica.

En el Ejemplo 2.2 calculamos los cuantiles al 20 %, 50 % y 80 % de este conjunto de datos. Para encontrar nuevamente estas cantidades usando ahora la función de distribución empírica simplemente marcamos los valores 0.2,

0.5 y 0.8 en el eje vertical como se muestra en la Figura 3.22. Después se traza una línea horizontal buscando la gráfica de $F(x)$ (considerando la línea punteada como parte de la gráfica) y al encontrarla se continúa con una línea vertical hacia abajo hasta alcanzar el eje horizontal. El valor x así encontrado es el cuantil correspondiente. Si el nivel buscado coincide con el piso de un escalón (esto ocurre en el caso del cuantil al 50 % en este ejemplo), la línea vertical se traza desde en el punto central del intervalo.

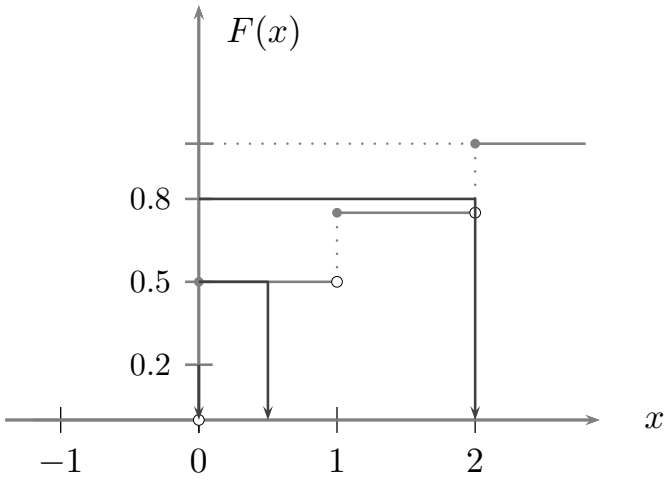


Figura 3.22: Procedimiento gráfico para el cálculo de cuantiles.

De esta manera gráfica se encuentra que $c_{0.2} = 0$, $c_{0.5} = 1/2$ (mediana) y $c_{0.8} = 2$. •

Ejercicios

79. Consulte un periódico y localice las posibles gráficas que allí aparezcan. Determine si algunas de estas gráficas corresponden o son similares a las que en este texto hemos presentado.

Gráfica de barras

80. De un grupo de 50 personas, 18 son fumadoras y 32 no son fumadoras. Elabore un gráfica de barras con esta información.

81. Acerca de la religión que profesa un grupo de 100 personas, se obtuvo la información que se presenta en la siguiente tabla. Elabore un gráfica de barras con esta información.

Religión

Valor	Frecuencia
Ninguno	28
Catolicismo	21
Cristianismo	19
Islam	15
Budismo	10
Otra	7

82. En la siguiente tabla se muestran los destinos principales de 73 personas que salen del país. Elabore un diagrama de barras con esta información.

Destino	Frecuencia
Canada	12
Francia	15
Estados Unidos	25
Inglaterra	14
Alemania	7

83. En la tabla que aparece a continuación se muestran los 10 principales países productores de café en el mundo para el año 2013, según datos de la FAO. Elabore una gráfica de barras horizontal con esta información.

Principales países productores de café en 2013

País	Producción (en toneladas)
Brasil	2,964,538
Vietnam	1,326,688
Indonesia	675,800
Colombia	653,160
Etiopía	392,006
India	318,200
Honduras	280,697
Perú	256,241
Guatemala	248,668
México	231,596

84. Investigue la densidad poblacional por continente y elabore una gráfica de barras con los datos encontrados.

85. Investigue la densidad poblacional de los cinco países más densamente poblados y elabore una gráfica de barras con los datos encontrados.

86. Investigue la extensión territorial de los diez países más grandes en territorio y elabore una gráfica de barras horizontal con los datos encontrados.

Histograma

87. ¿Cuál es la diferencia entre un histograma y una gráfica de barras?

88. En la siguiente tabla se muestra el número de desperfectos que tiene cada uno de 105 productos examinados. Elabore un histograma con esta información.

Números de desperfectos por producto	Frecuencia
0	82
1	15
2	5
3	2
4	1

89. De un grupo de 60 familias, se obtuvieron los datos que aparecen en la siguiente tabla acerca del número de automóviles por familia. Elabore un histograma con esta información.

Número de automóviles por familia					
Valor	0	1	2	3	4
Frecuencia	26	20	10	3	1

90. Se consultaron a 73 personas y se les preguntó por el número de visitas promedio al dentista al año, obteniendo los datos que aparecen en la siguiente tabla. Elabore un histograma con esta información.

Número de visitas al dentista							
Valor	0	1	2	3	4	5	6
Frecuencia	32	1	4	5	10	16	5

91. El número de días consecutivos que un grupo de trabajadores no pudo asistir a laborar por enfermedad tiene la frecuencia que se muestra en la siguiente tabla. Elabore un histograma con esta información.

Número de días con falta al trabajo					
Valor	1	2	3	4	5
Frecuencia	23	12	5	3	1

Gráfica de pastel

92. Elabore una gráfica de pastel para los datos que se muestran en la siguiente tabla para la variable número de hijos por familia. En total se consultaron a 120 familias.

Número de hijos por familia				
Valor	0	1	2	3
Frecuencia	24	78	12	6

93. Elabore un gráfica de pastel para los datos que muestran a continuación relativos a la composición de un país por clase socioeconómica.

Clase socioeconómica			
Valor	Baja	Media	Alta
Porcentaje	50 %	35 %	15 %

94. Elabore un gráfica de pastel para los datos que se muestran en la siguiente tabla para la variable número de padres vivos de una persona, en conjunto de 60 personas.

Número de padres vivos			
Valor	0	1	2
Frecuencia	5	10	45

95. Elabore una gráfica de pastel para los datos que se muestran a continuación de la variable número de goles anotados por un equipo de fútbol por partido jugado.

Número de goles anotados por partido							
Valor	0	1	2	3	4	5	6
Porcentaje	31 %	40 %	20 %	4 %	3 %	1 %	1 %

96. Hasta el año 2017, el número de campeonatos mundiales de fútbol ganados por país se muestra en la siguiente tabla. Elabore una gráfica de pastel con esta información. Si le es posible actualice la información a la fecha actual.

País	Campeonatos
Brasil	5
Alemania	4
Italia	4
Argentina	2
Uruguay	2
Francia	1
Inglaterra	1
España	1

Gráfica de tallo y hojas

97. Elabore un diagrama de tallo y hojas a partir del siguiente conjunto de datos indicando la unidad de la hoja.

a)

49	33	40	37	56	44	46	57	55	32
50	52	43	64	40	46	24	30	37	43
31	43	50	36	61	27	44	35	31	43
52	43	66	50	31	72	26	59	21	47

b)

1266	1354	1402	1107	1296	1389	1425
1087	1534	1200	1438	1024	1054	1190
1271	1342	1402	1055	1220	1372	1510
1124	1050	1199	1203	1355	1510	1426

c)

25.3	28.2	31.4	27.1	30.4	25.0
23.9	24.5	23.1	29.4	28.2	28.1
27.4	26.8	25.2	30.5	29.7	28.4
31.7	29.3	28.5	29.8	30.2	27.6

Diagrama de caja y brazos

98. Usando como parámetros la media, la desviación estándar y los valores máximo y mínimo, construya un diagrama de caja y brazos para el siguiente conjunto de datos.

    a) $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$.

    b) $-20, -1, -1, 0, 0, 1, 1, 15$.

    c) $2, 20, 4, 30, 5, 0, 10, 20$.

99. Usando como parámetros los cuantiles y $\pm(1.5)$ veces el rango intercuartil RIC$=Q_{0.75} - Q_{0.25}$, construya un diagrama de caja y brazos para el siguiente conjunto de datos.

    a) $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$.

    b) $-20, -1, -1, 0, 0, 1, 1, 15$.

    c) $2, 20, 4, 30, 5, 0, 10, 20$.

Función de distribución empírica

100. Encuentre la expresión de la función de distribución empírica del siguiente conjunto de datos. Grafique además esta función.

    a) $2, 5$.

    b) $-1, 0, 1$.

    c) $2, 0, 0, 1, 5, 3$.

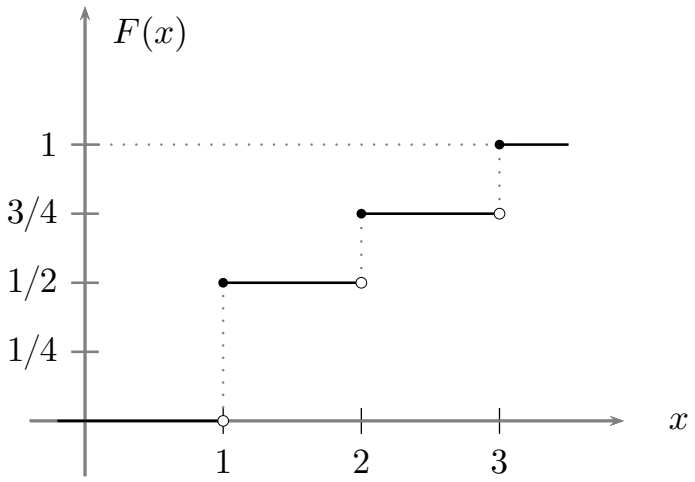
    d) $4, 10, 10, 4, 10, 4$.

    e) 7. (Un solo dato)

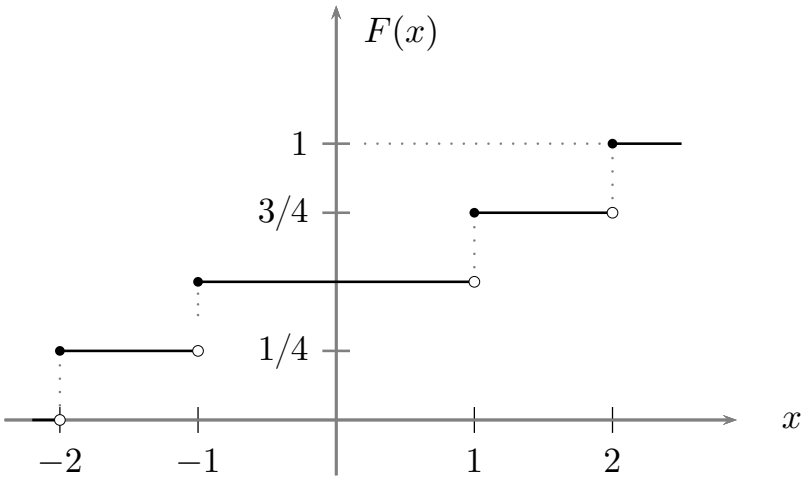
    f) $25, 25, 25, 25$.

101. Un cierto conjunto de datos $x_1, \ldots, x_n$ produce la función de distribución empírica $F(x)$ que aparece en cada uno de los siguientes incisos. Encuentre explícitamente a este conjunto de datos y escriba la expresión analítica de la función $F(x)$.

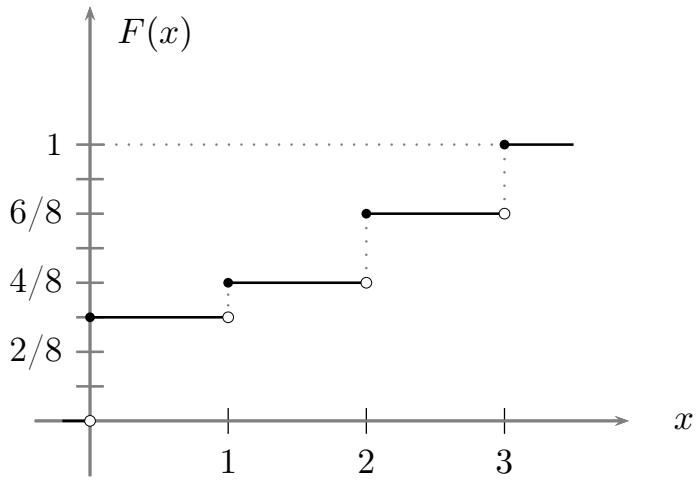
a)



b)



c)



102. Calcule los cuantiles al 20 %, 40 %, 60 % y 80 % del conjunto de datos resumido en la gráfica de la función de distribución empírica que se encuentra en la Figura 3.23.

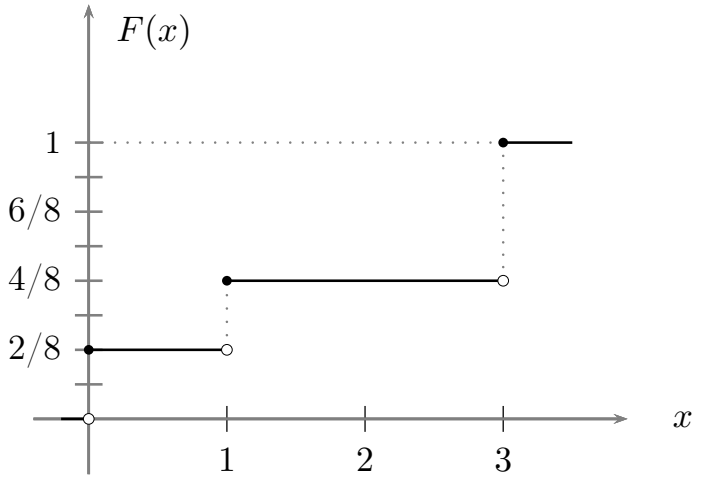


Figura 3.23: Una función de distribución empírica.

Capítulo 4

Descripciones para datos conjuntos

En este capítulo consideraremos que tenemos información simultánea de dos variables, es decir, las observaciones constan de una sucesión de parejas de datos

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Esta situación corresponde al hecho de llevar a cabo dos mediciones o dos preguntas a cada elemento de una muestra de tamaño n. Por ejemplo, para una población humana podemos considerar las variables edad y estatura, y obtener datos como los siguientes:

Dato	(Edad, Estatura)
1	$(20, 1.65)$
2	$(25, 1.70)$
3	$(20, 1.72)$
4	$(21, 1.75)$
5	$(23, 1.80)$
6	$(21, 1.68)$

Como antes, cada una de estas variables puede ser cualitativa o cuantitativa. Para el primer caso, puede ser además nominal u ordinal, y para el segundo

caso puede ser discreta o continua. Consideraremos principalmente el caso cuando las variables son cuantitativas.

Retomando el ejemplo mencionado, muchas otras mediciones o preguntas pueden solicitarse de una misma persona. Si incrementamos el número de preguntas, obtendríamos arreglos lineales (vectores) que contienen las respuestas a las preguntas planteadas. En efecto, si aplicamos un cuestionario de k preguntas a una persona, obtenemos como respuesta un vector $(x, y, \ldots)$ de k entradas. Cada una de estas entradas corresponde a las respuestas obtenidas.

En este capítulo definiremos algunas descripciones numéricas y gráficas en el caso de considerar dos variables a la vez. En particular, estaremos interesados en detectar alguna posible relación estadística entre dos variables. Gráficamente los datos $(x_1, y_1), \ldots, (x_n, y_n)$ pueden representarse como puntos o cruces en un plano como el que se muestra en la Figura 4.1. A este tipo de gráficas se les llama **diagramas de dispersión**.

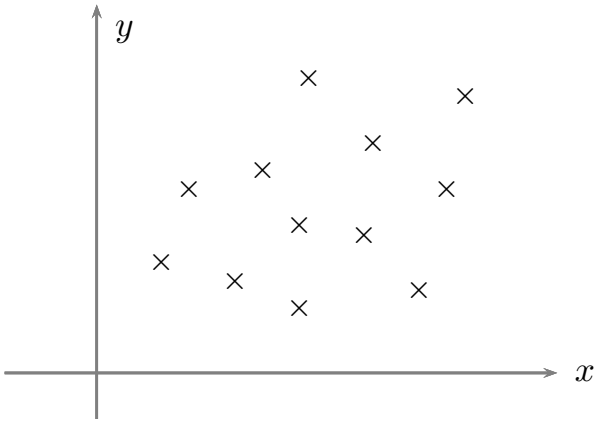


Figura 4.1: Ejemplo de diagrama de dispersión.

Los diagramas de dispersión ayudan a identificar visualmente alguna posible relación entre las dos variables en estudio. En general, pueden existir varios tipos de relaciones entre dos variables pero distinguiremos dos casos: relaciones lineales y relaciones no lineales. En el primer caso el diagrama de

dispersión presenta un cúmulo alargado de puntos que sugiere la relación lineal entre las variables, véanse las dos gráficas de la Figura 4.2, por ejemplo. En el segundo caso el diagrama de dispersión puede presentar un cúmulo de puntos con alguna tendencia curva o ninguna tendencia en lo absoluto. Particularmente estaremos interesados en detectar tendencias de relación lineal entre las variables. Diremos entonces que:

1. Existe **correlación positiva** entre las variables cuando los valores de y tienden a crecer de manera lineal conforme los valores de x crecen. Este comportamiento se muestra en la gráfica de la izquierda de la Figura 4.2.

2. Existe **correlación negativa** entre las variables cuando los valores de y tienden a decrecer de manera lineal conforme los valores de x crecen. Este comportamiento se muestra en la gráfica de la derecha de la Figura 4.2.

3. No existe **correlación** entre las variables cuando ninguna de las dos tendencias anteriores se presenta.

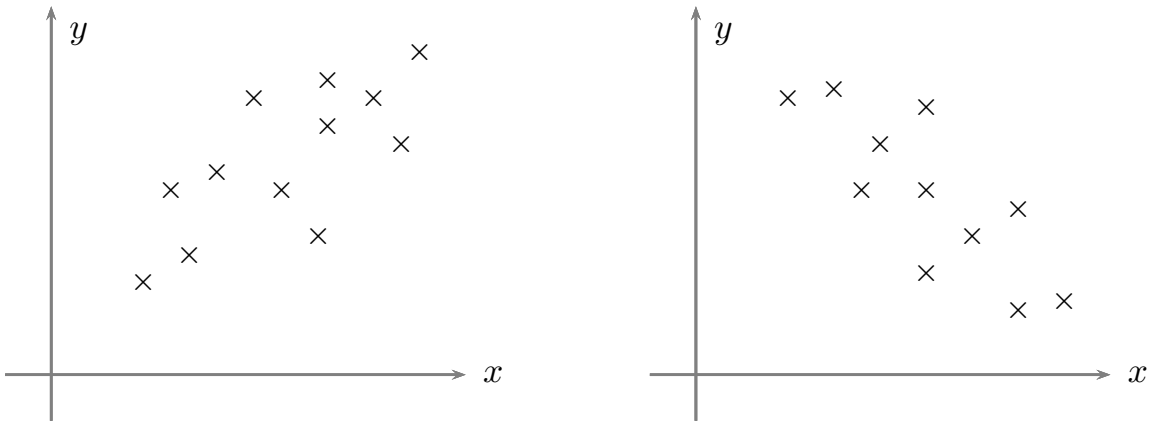


Figura 4.2: Correlación positiva (gráfica izquierda) y correlación negativa (gráfica derecha).

En R pueden usarse los siguientes comandos para visualizar un diagrama de

dispersión.

```
> x <- c(3,1.5,2.5,7,4,8,5.5,6,2,3.5,4,5,5)
> y <- c(3.5,3,5,4.5,3,2,5,3,4,2.5,4,2.5,3.5)
> plot(x,y)
```

El resultado es como se muestra en la Figura 4.3.

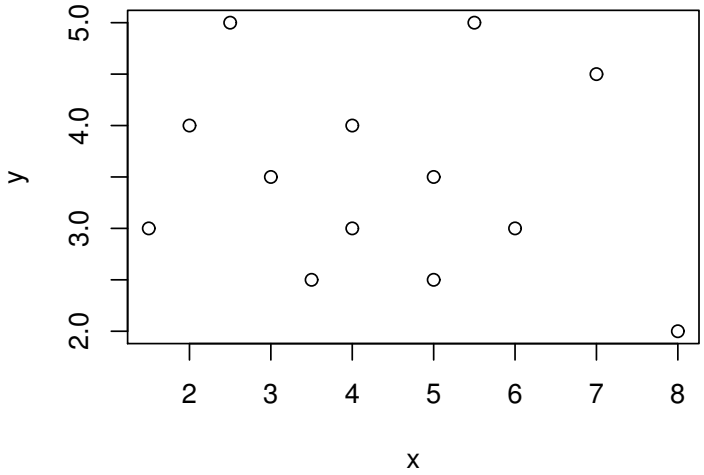


Figura 4.3: Diagrama de dispersión producido en el paquete R.

Ahora podemos empezar a cuantificar lo mencionado antes, no sin antes reiterar que pueden existir varios tipos de relaciones entre dos variables y que las cantidades que estudiaremos se refieren únicamente a las relaciones de tipo lineal.

Frecuencias para datos conjuntos

Cuando la colección de datos conjuntos $(x_1, y_1), \ldots, (x_n, y_n)$ es muy grande, es posible que sea conveniente agrupar esta información en un arreglo rectangular llamado tabla de frecuencias o tabla de contingencias. También es posible que la información se encuentre ya almacenada en este formato. Esta tabla es simplemente un arreglo como el que aparece a continuación.

$x \setminus y$	-2	-1	0	1	2
1	3	1	4	4	0
2	1	2	7	2	3
3	0	1	5	3	4
4	4	8	1	4	2
5	2	0	3	5	1

La primera columna de esta tabla corresponde a los valores tomados por la primera variable y el primer renglón son los valores de la segunda variable. Las entradas de esta tabla son las frecuencias con las que se observaron cada una de las parejas de valores. Por ejemplo, la pareja $(1, -2)$ se observó 3 veces. A partir de la información de esta tabla se pueden calcular las características numéricas de cada una de las variables por separado. Por ejemplo, la tabla de frecuencias de los valores de la primera variable se obtienen sumando todas las entradas de un mismo renglón, esto es,

x	Frecuencia
1	12
2	15
3	13
4	19
5	11

Análogamente, las frecuencias de los valores de la segunda variable se obtienen sumando las correspondientes columnas de la tabla conjunta.

Se pueden elaborar también gráficas de las frecuencias de datos conjuntos como la que se muestra en la Figura 4.4. En cada punto (x_i, y_j) se ha colocado un círculo de radio proporcional a la frecuencia en ese punto.

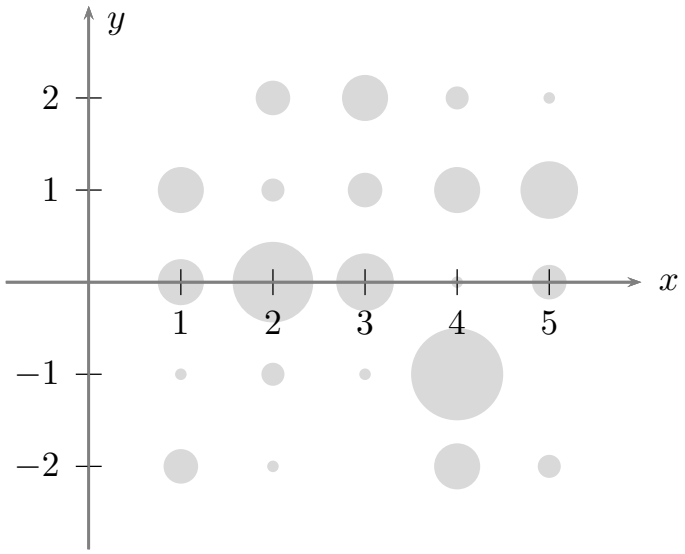


Figura 4.4: Una representación gráfica de frecuencias para datos conjuntos.

En este ejemplo, puesto que la frecuencia máxima es 8 y para que el radio máximo fuera de 0.5 se tomó cada frecuencia y se dividió entre 16. Otras figuras geométricas planas o en tres dimensiones pueden colocarse. Este tipo de gráficas son generalizaciones de las gráficas de barras o de frecuencias que hemos mencionado antes. Para mayor precisión, y si no resulta demasiada información aglomerada, se puede colocar el valor de la frecuencia en el centro de cada círculo.

Covarianza

Sea $(x_1, y_1), \ldots, (x_n, y_n)$ una colección de datos numéricos conjuntos de dos variables cuantitativas. Sea $\bar{x}$ la media de la primera variable y sea $\bar{y}$ la media de la segunda variable. La covarianza entre estas dos variables es un número que se denota por $\text{cov}(x, y)$ y se calcula como sigue.

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

La covarianza puede ser positiva, negativa o nula. Como veremos en la siguiente sección, la covarianza está relacionada con la correlación lineal entre dos variables. Cuando la covarianza es positiva, indica que existe algún grado de correlación lineal positiva (y crece cuando x crece) entre las variables. Cuando es negativa, indica que existe algún grado de correlación lineal negativa (y decrece cuando x crece) entre las variables. Cuando la covarianza es cero, indica la ausencia de dependencia lineal entre las variables. En este último caso, pudiera existir una relación no lineal perfecta entre las dos variables pero la covarianza no la detecta.

En el paquete `R`, el cálculo de la covarianza se efectúa mediante la función `cov()`. El siguiente es un ejemplo y es necesario advertir que `R` utiliza el valor $n - 1$ como denominador en la fórmula para la covarianza.


```
> x <- c(0,1,2)
> y <- c(3,2,4)
> cov(x,y)
[1] 0.5
```

Veamos ahora algunas propiedades de la covarianza.

- Las unidades de medición de la covarianza son las unidades de medición de la primera variable multiplicadas por las unidades de medición de la segunda variable. De esta manera, la covarianza está definida en términos de este producto de unidades de medición.

- El término covarianza sugiere alguna relación con la varianza vista antes. Efectivamente, tal relación existe: si calculamos la covarianza de las observaciones $x_1, \ldots, x_n$ de una variable x consigo misma obtene-

mos la varianza, pues

$$\begin{aligned}\operatorname{cov}(x,x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})\\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\\ &= \operatorname{var}(x).\end{aligned}$$

Puesto que la varianza es una cantidad mayor o igual a cero, esto indica que existe una correlación lineal positiva entre una variable consigo misma. Esto es evidente, la correlación lineal positiva es claramente perfecta en este caso.

- Debido a que el orden de los factores no altera el producto de cualesquiera dos números reales, tenemos que la covarianza es simétrica en sus argumentos, esto es,

$$\operatorname{cov}(x,y) = \operatorname{cov}(y,x).$$

Observe que la posible correlación lineal, positiva o negativa, se conserva al graficar los datos en el orden (x,y) o en el orden (y,x). Para experimentar esto trace usted una línea recta de pendiente positiva o negativa en un plano coordenado sobre una hoja de papel. Ahora vea el reverso de la hoja con el eje horizontal visto como el eje vertical. ¿Se preserva el tipo de relación lineal?

- Sea a una constante y denotemos por ax el conjunto de datos transformados $ax_1, \ldots, ax_n$. Esto corresponde a llevar a cabo un cambio de escala en esta variable. Recordemos que la media de estos nuevos datos es $a\bar{x}$. Entonces

$$\begin{aligned}\operatorname{cov}(ax,y) &= \frac{1}{n}\sum_{i=1}^{n}(ax_i-a\bar{x})(y_i-\bar{y})\\ &= \frac{1}{n}\sum_{i=1}^{n}a(x_i-\bar{x})(y_i-\bar{y})\\ &= a\,\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\\ &= a\cdot\operatorname{cov}(x,y).\end{aligned}$$

Esto significa que la covarianza hereda de manera idéntica un cambio de escala en cualquiera de las variables.

- Sea c una constante y denotemos por $x + c$ el conjunto de datos trasladados $x_1 + c, \ldots, x_n + c$. Recordemos que la media de estos nuevos datos es $\bar{x} + c$. Entonces

$$\begin{aligned} \operatorname{cov}(x + c, y) &= \frac{1}{n} \sum_{i=1}^{n} ((x_i + c) - (\bar{x} + c))(y_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\ &= \operatorname{cov}(x, y). \end{aligned}$$

  Esto significa que trasladar las observaciones de una o de otra variable, no tiene ningún efecto sobre la covarianza: su posible correlación lineal positiva o negativa sigue siendo la misma.

- Finalmente, encontraremos una fórmula alternativa para el cálculo de la covarianza. Multiplicando término a término la expresión de cada sumando en la definición de covarianza, tenemos que

$$\begin{aligned} \operatorname{cov}(x, y) &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \bar{y} - \frac{1}{n} \sum_{i=1}^{n} \bar{x} y_i + \frac{1}{n} \sum_{i=1}^{n} \bar{x}\bar{y} \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{y} \frac{1}{n} \sum_{i=1}^{n} x_i - \bar{x} \frac{1}{n} \sum_{i=1}^{n} y_i + \bar{x}\bar{y} \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{y}\,\bar{x} - \bar{x}\,\bar{y} + \bar{x}\bar{y} \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}. \end{aligned}$$

Por lo tanto, la covarianza puede calcularse a partir de la media de cada una de las variables y la suma indicada como aparece en la última expresión.

Otras propiedades sencillas de la covarianza se encuentran en la sección de ejercicios. En el siguiente recuadro se resumen algunas de las propiedades de la covarianza.

Propiedades de la covarianza

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- $\text{cov}(x, y) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$
- $\text{cov}(x, y) = \text{cov}(y, x)$
- $\text{cov}(x, x) = \text{var}(x)$
- $\text{cov}(x, c) = 0, \quad c$ constante
- $\text{cov}(ax, y) = a \cdot \text{cov}(x, y), \quad a$ constante
- $\text{cov}(x + c, y) = \text{cov}(x, y), \quad c$ constante

Coeficiente de correlación

Una medida entre dos variables que se encuentra estrechamente relacionada con la covarianza es el así llamado coeficiente de correlación. Consideremos nuevamente las observaciones $(x_1, y_1), \ldots, (x_n, y_n)$ de datos numéricos conjuntos de dos variables cuantitativas. Sea $\text{cov}(x, y)$ la covarianza entre estas variables como se ha definido antes y sean $\text{var}(x)$ y $\text{var}(y)$ las correspondientes varianzas. El coeficiente de correlación entre estas dos variables es un número que se denota por $\rho(x, y)$, en donde la letra ρ pertenece al alfabeto

griego y se le llama ro. Este coeficiente se define de la siguiente forma.

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}}$$

En la expresión anterior suponemos que las varianzas indicadas son distintas de cero para que el cociente esté bien definido. En el paquete `R`, el cálculo del coeficiente de correlación se efectúa mediante la función `cor()`. El siguiente es un ejemplo.

R

```
> x <- c(0,1,2)
> y <- c(3,0,4)
> cor(x,y)
[1] 0.2401922
```

Hemos advertido antes que `R` utiliza el número $n - 1$ como denominador en las fórmulas para la covarianza y la varianza. Dado que $\rho(x, y)$ es el cociente arriba indicado, puede verificarse que esta diferencia no afecta el valor del coeficiente.

El coeficiente de correlación es la medida comúnmente usada para determinar el grado de dependencia lineal entre dos variables. Esto se debe a que el coeficiente de correlación cumple las siguientes tres propiedades importantes:

1. El coeficiente de correlación toma siempre valores reales entre -1 y $+1$ inclusive, es decir, para cualesquiera observaciones $(x_1, y_1), \ldots, (x_n, y_n)$, se cumple que
$$-1 \leqslant \rho(x, y) \leqslant 1.$$

2. El coeficiente de correlación toma el valor $+1$ si, y sólo si, existe una correlación positiva perfecta entre las variables, es decir, existen constante a y b, con $a > 0$, tales que $y = ax + b$.

3. El coeficiente de correlación toma el valor -1 si, y sólo si, existe una correlación negativa perfecta entre las variables, es decir, existen constantes a y b, con $a < 0$, tales que $y = ax + b$.

No presentaremos las demostraciones de estos resultados pues se requiere un tratamiento matemático mayor al supuesto para el presente trabajo. Sin embargo, vamos a hacer algunas observaciones generales para tratar de entender mejor estas propiedades.

- Nuestra primera observación es que el coeficiente de correlación no posee unidad de medición pues las unidades de medición de las variables involucradas se pierden al llevar a cabo el cociente indicado. Esta propiedad de falta de unidad de medición es buena pues de esta manera se pueden hacer comparaciones entre los valores de este coeficiente sin importar la naturaleza de las variables en estudio.

- Debido a que el orden de los factores no altera el producto de cualesquiera dos números reales, tenemos que el coeficiente de correlación, como la covarianza, es simétrico en sus argumentos, esto es,

$$\rho(x, y) = \rho(y, x).$$

- Sea a una constante distinta de cero y denotemos por ax el conjunto de datos transformados $ax_1, \ldots, ax_n$. Esto corresponde a llevar a cabo un cambio de escala en esta variable. Recordemos que la media de estos nuevos datos es $a\bar{x}$. Entonces

$$\begin{aligned}
\rho(ax, y) &= \frac{\text{cov}(ax, y)}{\sqrt{\text{var}(ax) \cdot \text{var}(y)}} \\
&= \frac{a \cdot \text{cov}(x, y)}{\sqrt{a^2 \cdot \text{var}(x) \cdot \text{var}(y)}} \\
&= \frac{a}{|a|} \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}} \\
&= \frac{a}{|a|} \rho(x, y) \\
&= \begin{cases} +\rho(x, y) & \text{si } \ a > 0, \\ -\rho(x, y) & \text{si } \ a < 0. \end{cases}
\end{aligned}$$

  Esto significa que la magnitud del coeficiente de correlación se preserva bajo la transformación indicada, pero adquiere el signo de la constante a.

- Sea c una constante y denotemos por $x + c$ el conjunto de datos trasladados $x_1 + c, \ldots, x_n + c$. Recordemos que la media de estos nuevos datos es $\bar{x} + c$. Entonces

$$\begin{aligned} \rho(x + c, y) &= \frac{\text{cov}(x + c, y)}{\sqrt{\text{var}(x + c) \cdot \text{var}(y)}} \\ &= \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}} \\ &= \rho(x, y). \end{aligned}$$

  Esto significa que trasladar las observaciones de una o de otra variable, no tiene ningún efecto sobre el coeficiente de correlación.

Finalmente comentaremos la propiedad más importante del coeficiente de correlación: aquella que establece que es una medida del grado de dependencia lineal entre las variables.

- Mientras más cercano sea el coeficiente de correlación al valor 1 o al valor -1, mayor dependencia lineal existe entre las dos variables. Consideremos el caso extremo cuando las observaciones de la segunda variable están dadas por $y_i = ax_i + b$, con $a \neq 0$ y b dos constantes. Entonces, haciendo uso de las propiedades de la covarianza y la varianza, tenemos que

$$\begin{aligned} \rho(x, ax + b) &= \frac{\text{cov}(x, ax + b)}{\sqrt{\text{var}(x) \cdot \text{var}(ax + b)}} \\ &= \frac{a \cdot \text{cov}(x, x)}{\sqrt{\text{var}(x) \cdot a^2 \cdot \text{var}(x)}} \\ &= \frac{a \cdot \text{var}(x)}{|a| \cdot \text{var}(x)} \\ &= \frac{a}{|a|} \\ &= \begin{cases} +1 & \text{si } a > 0, \\ -1 & \text{si } a < 0. \end{cases} \end{aligned}$$

  Cuando el coeficiente de correlación es cercano a cero, indica la ausencia de dependencia lineal entre las variables. Se debe advertir que

existen múltiples ejemplos de dos variables que evolucionan en el tiempo y que presentan un coeficiente de correlación alto, pero en realidad no hay ninguna o muy poca relación causal entre ellas, por ejemplo, el número de premios Nobel obtenidos por país y el consumo de chocolate per cápita. Véase el divertido libro de Vigen [18] en donde se muestran las gráficas de varios de estos ejemplos.

En el siguiente recuadro se resumen algunas de las propiedades del coeficiente de correlación.

Propiedades del coeficiente de correlación

$$\rho(x,y) = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)\cdot\operatorname{var}(y)}}$$

- $-1 \leqslant \rho(x,y) \leqslant 1$
- $\rho(x,y) = \rho(y,x)$
- $\rho(ax,y) = \begin{cases} +\rho(x,y) & \text{si } a>0, \\ -\rho(x,y) & \text{si } a<0. \end{cases}$
- $\rho(x+c,y) = \rho(x,y), \quad c$ constante
- $\rho(x,y) = 1 \Leftrightarrow y = ax+b$, con a,b constantes, a$>$0
- $\rho(x,y) = -1 \Leftrightarrow y = ax+b$, con a,b constantes, a$<$0

Recta de regresión

Una vez que uno ha determinado que existe una posible dependencia lineal y causal entre las observaciones de dos variables, el siguiente paso consiste en encontrar la línea recta que mejor se adapte a los datos. Mediante el así llamado método de mínimos cuadrados, puede comprobarse que esta recta tiene la siguiente fórmula:

$$y - \bar{y} = \frac{\text{cov}(x, y)}{\text{var}(x)} (x - \bar{x}),$$

en donde las cantidades $\bar{x}$, $\bar{y}$, $\text{var}(x)$ y $\text{cov}(x, y)$ son constantes y se calculan como hemos indicado antes usando las observaciones $(x_1, y_1), \ldots, (x_n, y_n)$. A esta línea se le llama recta de regresión de la variable y sobre la variable x. Véase la Figura 4.5 en donde se muestra un diagrama de dispersión y la recta de regresión correspondiente. De la ecuación anterior puede comprobarse inmediatamente que la recta de regresión pasa por el punto $(\bar{x}, \bar{y})$.

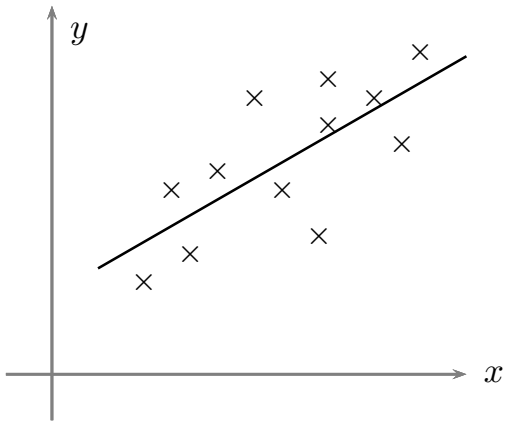


Figura 4.5: Ejemplo de recta de regresión.

Debemos observar que siempre se puede construir la recta de regresión para una colección dada de observaciones $(x_1, y_1), \ldots, (x_n, y_n)$. Sin embargo, esta recta, como modelo de dependencia lineal entre las variables x y y, sólo tendrá sentido cuando el coeficiente de correlación así lo indique tomando un valor cercano a $+1$ o -1. Suponiendo que tal es el caso y dado un valor x_0 para x, se puede estimar el valor de y por el valor

$$y_0 = \bar{y} + \frac{\text{cov}(x, y)}{\text{var}(x)}(x_0 - \bar{x}).$$

Generalmente estas estimaciones son razonables si el punto dado x_0 se en-

cuentra dentro del rango o intervalo de valores observados para x. Al proceso de estimar el valor de y para algún valor x_0 fuera del intervalo de observaciones para x se le llama extrapolación, pero no debe confiarse demasiado en tales estimaciones pues el modelo lineal puede no ser válido fuera de la ventana determinada por las observaciones. De la misma manera y bajo las precauciones debidas, puede estimarse el valor de x para un valor dado de y.

En el paquete R puede obtenerse de manera gráfica la recta de regresión usando las funciones que se muestran en el siguiente recuadro.


```
> x <- c(0,1,2,3,4)
> y <- c(2,4,5,7,6)
> plot(x,y)
> abline(lm(y~x))
```

El resultado se muestra en la Figura 4.6.

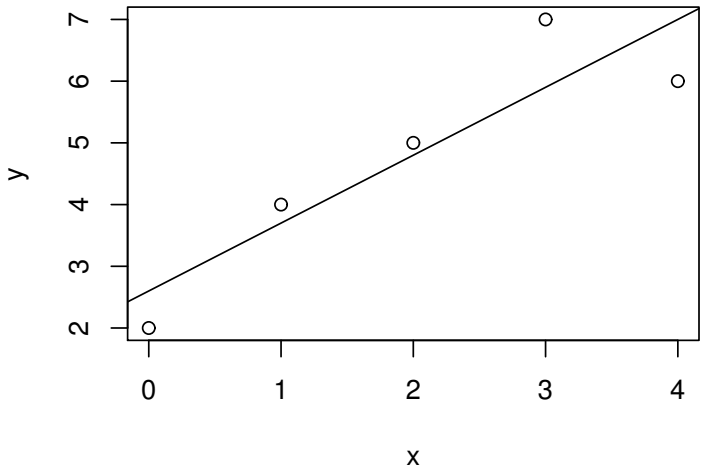


Figura 4.6: Gráfica de una recta de regresión producida en el paquete R.

Gráfica Q-Q

Sean $x_1, \ldots, x_n$ y $y_1, \ldots, y_m$ dos conjuntos de datos numéricos, no necesariamente del mismo tamaño. Una gráfica Q-Q, también llamada gráfica cuantil-

cuantil, permite determinar visualmente si estos dos conjuntos de datos son observaciones de dos variables con la misma distribución de probabilidad, y en tal caso deben tener aproximadamente las mismas características estadísticas.

Una gráfica Q-Q está basada en el cálculo de los cuantiles de los dos conjuntos de datos. Explicaremos a continuación su construcción. Denotemos por las siguientes expresiones los 99 percentiles del primer conjunto de datos

$$Q^x_{0.01}, Q^x_{0.02}, \ldots, Q^x_{0.99}.$$

Observe que varios de estos percentiles son idénticos si el número de datos $x_1, \ldots, x_n$ es pequeño, o bien si en esta lista aparecen muchos datos repetidos. Consideraremos también los percentiles para el segundo conjunto de datos

$$Q^y_{0.01}, Q^y_{0.02}, \ldots, Q^y_{0.99}.$$

Observe que en estas dos listas de percentiles no aparecen de manera explícita los tamaños de muestra n y m de los datos, y que estamos considerando la totalidad de los 99 percentiles. En una gráfica Q-Q se identifican con una cruz (o un punto) en un plano coordenado los puntos con coordenadas

$$(Q^x_{0.01}, Q^y_{0.01}), (Q^x_{0.02}, Q^y_{0.02}), \ldots, (Q^x_{0.99}, Q^y_{0.99}),$$

y para fines de comparación se puede trazar al mismo tiempo la recta de la función identidad. Un ejemplo de esta gráfica se muestra en la Figura 4.7.

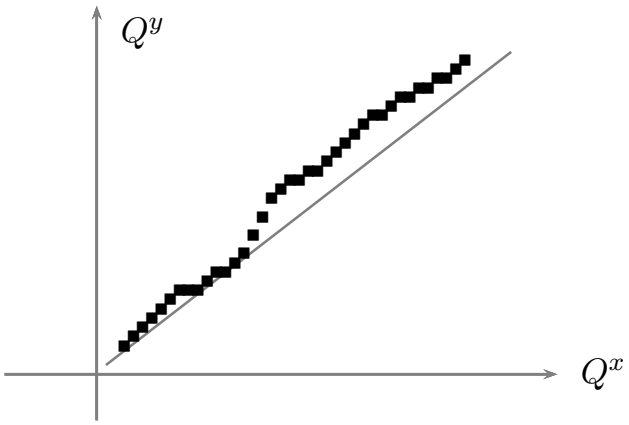


Figura 4.7: Ejemplo de una gráfica Q-Q.

Cuando los números de observaciones n y m son grandes y si, efectivamente, ambos conjuntos de observaciones provienen de variables con las mismas características estadísticas, la trayectoria de puntos dados por los cuantiles tiende a parecerse a la función identidad. En la explicación de la construcción de la gráfica Q-Q se usaron 99 cuantiles, pero un número mayor o menor de cuantiles puede ser usado.

En el paquete `R` se pueden elaborar gráficas Q-Q mediante el comando `qqplot()` como se muestra en el siguiente recuadro.

R

```
> x <- c(0,1,0,1,0,1)
> y <- c(1,0,1,0,1,0,1,0)
> qqplot(x,y)
```

En este ejemplo únicamente hay dos valores dentro de los conjuntos de datos x y y. Estos son los valores 0 y 1, y en ambos conjuntos de datos aparecen los valores 0 y 1 con la misma frecuencia. Podemos pensar que los valores 0 y 1 corresponden a las caras de una moneda y que las observaciones obtenidas pertenecen a resultados de lanzamientos de dos monedas equilibradas.

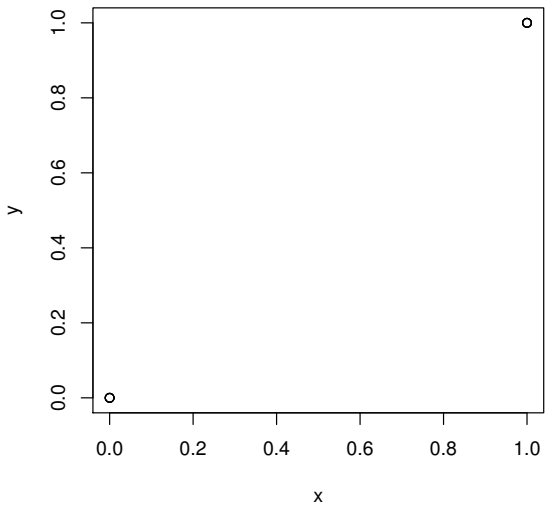


Figura 4.8: Ejemplo de una gráfica Q-Q producida en el paquete `R` usando el comando `qqplot()`.

El resultado que produce `R` con los comandos que aparecen en el recuadro anterior se muestra en la Figura 4.8. Observe que en este caso los cuantiles toman los valores 0 y 1, y en la gráfica aparecen únicamente dos puntos alineados sobre la recta de la función identidad. Esto se debe a que la frecuencia con la que aparecen los valores 0 y 1 en ambos conjuntos de datos son la misma, aunque los tamaños de los conjuntos sean distintos. La gráfica corrobora que los dos conjuntos de datos tienen las mismas características estadísticas.

Las gráficas Q-Q también se utilizan para determinar si un conjunto de observaciones provienen de una variable con un modelo (distribución de probabilidad) teórico dado. En este caso las observaciones se comparan con los cuantiles del modelo teórico. No abordaremos este tema aquí pues eso nos llevaría a revisar algunos modelos teóricos existentes, pero un caso importante es el de la distribución normal. Para llevar a cabo una comparación de un conjunto de datos respecto de la distribución normal estándar en el paquete `R` se usa el comando `qqnorm()`. Un ejemplo de esto se muestra en el siguiente recuadro y los resultados se muestran en la Figura 4.9.

R

```
> x <- c(-2,-1,0,0,0,1,2)
> qqnorm(x)
```

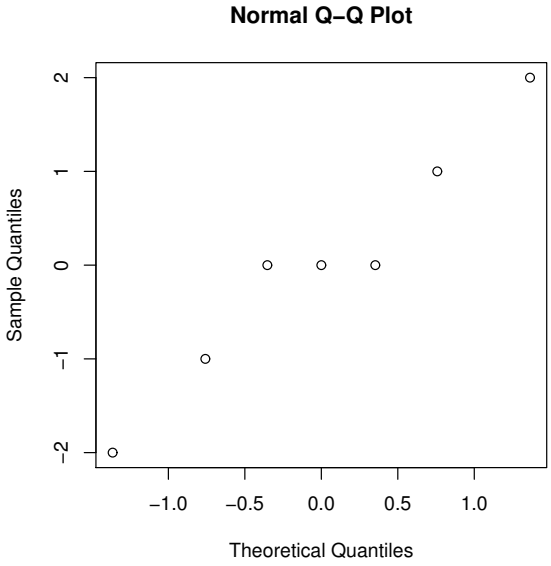


Figura 4.9: Ejemplo de una gráfica Q-Q producida en el paquete `R` usando el comando `qqnorm()`.

Para concluir este capítulo presentamos en la siguiente tabla un resumen de las fórmulas principales que hemos visto y que están relacionadas con la descripción de un conjunto de datos bivariados.

RESUMEN DE FÓRMULAS

Descripciones numéricas de un conjunto de datos bivariados $(x_1, y_1), \ldots, (x_n, y_n)$

Covarianza $$\operatorname{cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Coeficiente de correlación $$\rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x) \operatorname{var}(x)}}$$

Recta de regresión $$y - \bar{y} = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} (x - \bar{x})$$

Todas las descripciones numéricas individuales mencionadas antes pueden aplicarse a cada variable por separado para conformar vectores de descripciones numéricas. Por ejemplo, se puede considerar el vector de medias $(\bar{x}, \bar{y})$, el vector de modas, el vector de medianas, etc.

Ejercicios

103. Sin llevar a cabo ningún cálculo, determine si los conjuntos de puntos que aparecen en la Figura 4.10 presentan correlación positiva o negativa. Justifique su respuesta.

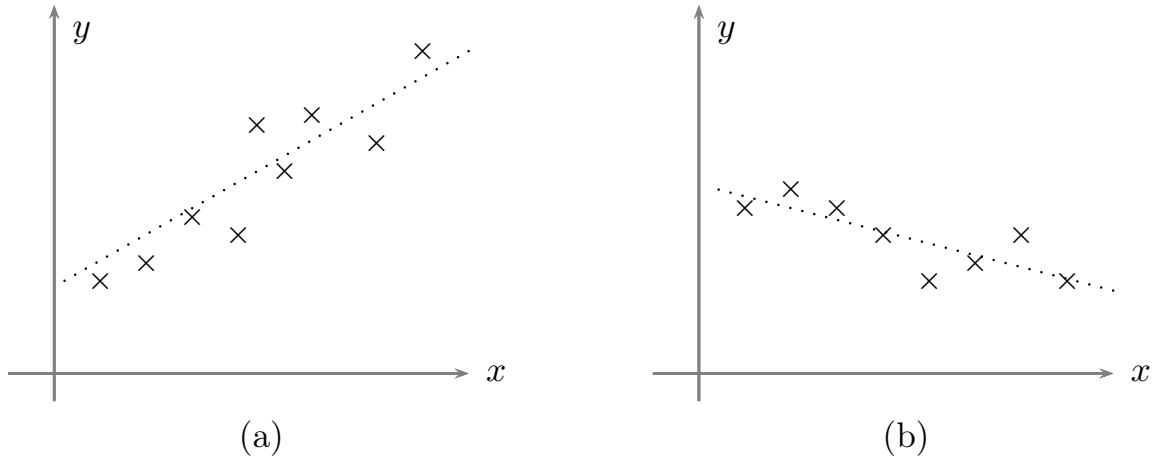


Figura 4.10

Frecuencias para datos conjuntos

104. Encuentre las frecuencias de cada una de las variables x y y por separado cuando la información conjunta está dada por la siguiente tabla de frecuencias.

$x \setminus y$	10	20	30	40
0	7	4	1	4
1	3	2	8	2
2	0	2	4	6
3	4	1	0	4

Covarianza

105. Calcule la covarianza del siguiente conjunto de datos.

a) $(0,0),(1,2),(2,1)$

b) $(-1,1),(0,-1)(1,-2)(2,-3)$.

c) $(2,4),(0,0),(1,2),(-1,-2),(0,0)$.

106. Sea $(x_1,y_1),\ldots,(x_n,y_n)$ una colección de datos numéricos conjuntos. Sean a,b,c y d cuatro constantes dadas. Denotemos por $ax+b$ al conjunto de datos transformados $ax_1+b,\ldots,ax_n+b$, y por $cy+d$ al conjunto de datos $cy_1+d,\ldots,cy_n+d$. Compruebe que

$$\operatorname{cov}(ax+b,cy+d)=ac\cdot\operatorname{cov}(x,y).$$

107. Sea c una constante y considere la colección de datos numéricos conjuntos $(x_1,c),\ldots,(x_n,c)$, en donde la segunda coordenada siempre es la constante c. Compruebe que

$$\operatorname{cov}(x,c)=0.$$

108. Sin llevar a cabo ningún cálculo, determine si los conjuntos de puntos que aparecen en la Figura 4.11 presentan covarianza grande o pequeña.

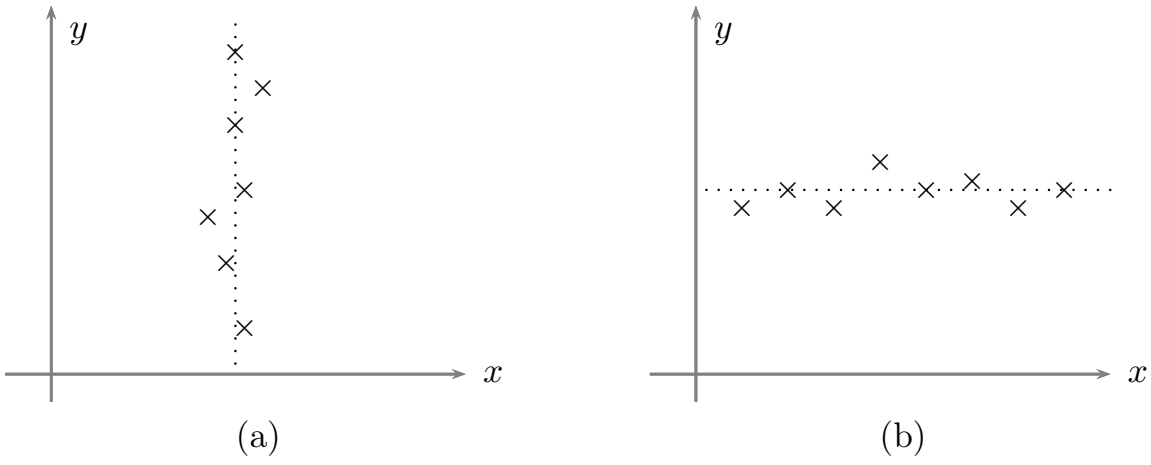


Figura 4.11

Coeficiente de correlación

109. Sea $(x_1,y_1),\ldots,(x_n,y_n)$ una colección de datos numéricos conjuntos. Sean a,b,c y d cuatro constantes dadas. Denotemos por $ax+b$ al

conjunto de datos transformados $ax_1 + b, \ldots, ax_n + b$, y por $cy + d$ al conjunto de datos $cy_1 + d, \ldots, cy_n + d$. Suponga que $a \neq 0$ y $c \neq 0$. Compruebe que

$$\rho(ax + b, cy + d) = \frac{a}{|a|}\frac{c}{|c|} \cdot \rho(x, y).$$

Recta de regresión

110. Calcule la recta de regresión para el conjunto de observaciones que aparecen en la siguiente tabla. Dibuje en un plano cartesiano el diagrama de dispersión de los datos, trace la recta de regresión y estime el valor de y cuando $x = 2.5$.

x	0	1	2	3	4
y	3.5	4.5	7.5	9.5	10

111. A través de una recta de regresión, estime el dato faltante en la siguiente tabla de registros.

x	1	2	3	4	5
y	4	8		18	22

112. Compruebe que el coeficiente principal de la recta de regresión se puede escribir de la siguiente forma

$$\rho(x, y) \cdot \sqrt{\frac{\text{var}(y)}{\text{var}(x)}}.$$

113. A partir de la ecuación de la recta de regresión construida usando el conjunto de observaciones $(x_1, y_1), \ldots, (x_n, y_n)$, compruebe que, dado

un valor y_0 dentro del rango de valores para y, el valor aproximado para x es x_0 dado por

$$x_0 = \bar{x} + \frac{\text{var}(x)}{\text{cov}(x,y)}(y_0 - \bar{y}).$$

114. Considere un conjunto de observaciones $(x_1, y_1), \ldots, (x_n, y_n)$ en donde cada coordenada se transforma en $u_i = ax_i + b$ y $v_i = cy_i + d$, con a, b, c y d constantes, $a \neq 0$, $c \neq 0$. Esta transformación corresponde a una traslación y cambio de escala en cada una de las variables. Compruebe que la recta de regresión lineal de las variables u y v está dada por

$$v = \frac{c}{a} \cdot \frac{\text{cov}(x,y)}{\text{var}(x)} (u - (a\bar{x} + b)) + (c\bar{y} + d).$$

Gráfica Q-Q

115. La colección de números x que aparece abajo son los resultados de 20 lanzamientos de un dado. La colección de números y corresponde al resultado teórico de obtener cada una de las caras de un dado equilibrado en seis lanzamientos. Compare los cuantiles de estos dos conjuntos de datos elaborando una gráfica Q-Q en `R` para determinar cualitativamente si las observaciones registradas corresponden a un dado equilibrado.

x : 5 4 6 4 2 2 2 3 6 1 5 2 6 3 5 5 3 4 2 1
y : 1 2 3 4 5 6

116. Elabore una gráfica Q-Q en `R` para determinar subjetivamente si los siguientes dos conjuntos de datos tienen características estadísticas similares.

x : 5 2 8 3 3 6 1 10 4 6 4 1 5 3 3 6 5 5 9 6
y : 2 5 5 8 2 8 6 5 2 2 7 6 5 5 5 5 6 3 6 5 3 6 5 4 9

117. Elabore una gráfica Q-Q en `R` para determinar subjetivamente si el siguiente conjuntos de datos tiene características estadísticas similares a la distribución normal estándar.

-1.259	-0.760	-0.247	1.123	-0.856
0.611	-1.257	0.280	-0.901	-0.424
0.004	-0.654	1.216	0.621	-1.385
0.522	1.731	0.210	-1.667	-0.440

Notación y términos matemáticos

Sumas

La suma de los números $x_1, \ldots, x_n$ se puede escribir de manera abreviada como sigue:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

Se pueden verificar entonces las siguientes propiedades: si c es una constante, entonces

1. $\displaystyle\sum_{i=1}^{n} c = nc.$

2. $\displaystyle\sum_{i=1}^{n} c\, x_i = c \sum_{i=1}^{n} x_i.$

3. $\displaystyle\sum_{i=1}^{n} (x_i + c) = \sum_{i=1}^{n} x_i + nc.$

4. $\displaystyle\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

5. $\displaystyle\sum_{i=1}^{n} (x_i + c)^2 = (\sum_{i=1}^{n} x_i^2) + 2c\,(\sum_{i=1}^{n} x_i) + nc^2.$

Exponentes

El término x^n denota el número x multiplicado por sí mismo n veces. Al número n se le llama exponente y no necesariamente es un número entero. A continuación se recuerdan algunas reglas para el manejo de exponentes.

1. $x^1 = x$.
2. $x^0 = 1, \quad x \neq 0$.
3. $0^n = 0, \quad n \neq 0$.
4. 0^0 no se define.
5. $x^{-1} = \dfrac{1}{x}, \quad x \neq 0$.
6. $x^n \cdot x^m = x^{n+m}$.
7. $\dfrac{x^n}{x^m} = x^{n-m}$.
8. $(x^n)^m = x^{nm}$.
9. $(xy)^n = x^n\, y^n$.
10. $\left(\dfrac{x}{y}\right)^n = \dfrac{x^n}{y^n}$.
11. $x^{-n} = \dfrac{1}{x^n}, \quad x \neq 0$.
12. $x^{m/n} = \sqrt[n]{x^m}$.

Valor absoluto

El valor absoluto de un número x se denota por $|x|$ y se define como sigue:

$$|x| = \begin{cases} +x & \text{si} \quad x \geqslant 0, \\ -x & \text{si} \quad x < 0. \end{cases}$$

De esta manera, el valor absoluto es siempre un número mayor o igual a cero. Así, tenemos por ejemplo,

$$\begin{aligned} |-7| &= 7, \\ |2.5| &= 2.5, \\ |-4/3| &= 4/3, \\ |0| &= 0. \end{aligned}$$

En particular, observe que

1. $\sqrt{x^2} = |x|$.

2. $\dfrac{x}{|x|} = \dfrac{|x|}{x} = \begin{cases} +1 & \text{si } x > 0, \\ -1 & \text{si } x < 0. \end{cases}$

Intervalos

Sean $a \leqslant b$ dos números reales. Los intervalos que aparecen en el lado izquierdo de la siguiente lista se definen como los conjuntos de números que aparecen en el lado derecho:

1. $(a, b) = \{\, x : a < x < b \,\}$.

2. $[a, b) = \{\, x : a \leqslant x < b \,\}$.

3. $(a, b] = \{\, x : a < x \leqslant b \,\}$.

4. $[a, b] = \{\, x : a \leqslant x \leqslant b \,\}$.

5. $(-\infty, a) = \{\, x : x < a \,\}$.

6. $(-\infty, a] = \{\, x : x \leqslant a \,\}$.

7. $(a, \infty) = \{\, x : x > a \,\}$.

8. $[a, \infty) = \{\, x : x \geqslant a \,\}$.

Potencia de un binomio

Para cualesquiera dos números reales a y b tenemos que se cumplen las siguientes igualdades:

1. $(a+b)^2 = a^2 + 2ab + b^2$.
2. $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$.
3. $(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$.

Más generalmente, el teorema del binomio establece que para cualquier entero $n \geqslant 1$,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k,$$

en donde $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$ es el así llamado coeficiente binomial y $n! = n(n-1)\cdots 1$ es el factorial del número entero $n \geqslant 1$. Se define además $0! = 1$.

Breve orientación sobre R

El programa R es un paquete estadístico de distribución libre. Por sus muchas bondades, es cada vez de mayor uso en las universidades y la industria. Computacionalmente es muy poderoso y muy completo. Está disponible para los principales sistemas operativos como MacOSX, Windows y las varias distribuciones de Linux. El programa se puede obtener de la dirección

`http://www.r-project.org`

Para iniciarse en el aprendizaje de R pueden consultarse los varios manuales que han escrito algunas personas que usan R en sus actividades profesionales, y que de manera altruista distribuyen libremente en internet. Puede escribirse en un buscador de internet la frase **‘‘R manual’’** para obtener algunos ejemplos de manuales actualizados. También se pueden consultar libros publicados como [7] u [8], por ejemplo. Existe también la revista *The R Journal*, en donde se publican con regularidad artículos de interés para usuarios y desarrolladores de R.

Para una experiencia más amigable en el uso de R se recomienda instalar además un programa adicional que permita crear programas (sucesiones de comandos) en R dentro de un ambiente gráfico e integrado. A estos programas se les llama *Integrated Development Environment* (IDE) o *Graphical Unit Interface* (GUI). Como ejemplos de este tipo de programas (gratis en su versión básica) tenemos **RStudio**, **Eclipse**, **R Commander** o **RKWard**, entre otros muchos. De esta manera, el usuario trabaja sobre una interfaz gráfica agradable y bien organizada como medio de interacción, mientras que R funciona en el fondo.

Como otros programas de cómputo, el paquete R también puede ser usado

en línea a través de un navegador. Esto tiene la ventaja de que no se necesita instalar nada en el equipo de cómputo del usuario pues los cálculos se desarrollan en el servidor en donde se encuentra instalado este servicio. Esta forma de usar `R` puede presentar varias limitaciones pero puede ser un buen punto de inicio para aquellas personas con pocos conocimientos de cómputo y quienes desean experimentar con algunos comandos básicos de `R`. Para encontrar ejemplos de sitios activos que ofrecen usar `R` en línea, es suficiente escribir la frase ‘‘`R online`’’ en un navegador de internet.

Otros lenguajes o programas de cómputo, libres o de pago, que pueden usarse para llevar a cabo algún tipo de análisis estadístico, ya sea de manera directa o añadiendo alguna extensión son: Excel, Matlab, SPPP, PSSS, SAS, Sage, Octave, Mathematica, Minitab, Stata, Maple, Python, etc.

Ha sido inevitable hablar de algún sistema cómputo como apoyo en el estudio de la estadística descriptiva. En este trabajo hemos dado algunos ejemplos del uso básico de comandos en `R` para hacer cálculos o gráficas. Sin embargo, dada la rápida obsolescencia de los sistemas de cómputo, se ha decidido por no hacer demasiado énfasis a estas importantes herramientas computacionales y, por el contrario, nos concentramos en las fórmulas y resultados matemáticos. Éstos últimos tienen un tiempo de vigencia infinitamente mayor que cualquier sistema de cómputo avanzado. Así, la información indicada en esta pequeña orientación acerca del paquete `R`, y los recuadros presentados a lo largo del texto sobre el uso de `R`, deben tomarse considerando su muy limitada vigencia.

Descripciones numéricas para variables aleatorias

Esta sección tiene el objetivo importante de situar los temas que se estudian en este pequeño libro dentro del contexto general de la teoría matemática de la probabilidad y la estadística. Se desea además extender las definiciones de las descripciones numéricas vistas antes para conjuntos de datos numéricos, ahora para variables aleatorias. Esta sección puede servir también de orientación para aquellos estudiantes que continuarán sus estudios con cursos de teoría de la probabilidad o de estadística inferencial, y que el temario de alguna de sus asignaturas contempla a la estadística descriptiva como una colección de herramientas útiles que es adecuado conocer pero también como una introducción a estudios más avanzados de la estadística. Nuestra exposición será breve y el tratamiento será de un nivel matemático mayor al hasta ahora presentado, de modo que el lector no debe preocuparse demasiado si algunos pasajes no son del todo comprensibles.

El concepto importante es el de variable aleatoria. Una variable aleatoria (se abrevia v.a.) es una función X definida sobre una población y cuyos valores son números reales. Puede interpretarse esta función como una pregunta o una medición que se hace sobre cada elemento de la población. En la probabilidad y la estadística se estudian a las variables aleatorias pues representan las características de la población que deseamos conocer. Pero cuando no nos es posible tener la información completa de la población es que consideramos la idea de tomar sólo algunos elementos y hacer en ellos las mediciones. En este trabajo hemos supuesto tener una muestra (subconjunto) de la población y hacer la medición sobre estos elementos produciendo los resultados

$x_1, \ldots, x_n$. Véase la Figura A.1. Más generalmente, uno puede pensar que se toma un elemento al azar de la población (aquí radica la aleatoriedad) y se efectúa la medición produciendo un valor x. Debido al carácter aleatorio con el que fue escogido el elemento de la población es que se piensa que el valor x fue generado al azar y por ello la función X adquiere el nombre de variable aleatoria, pero vista como una función, no hay ninguna aleatoriedad en ella.

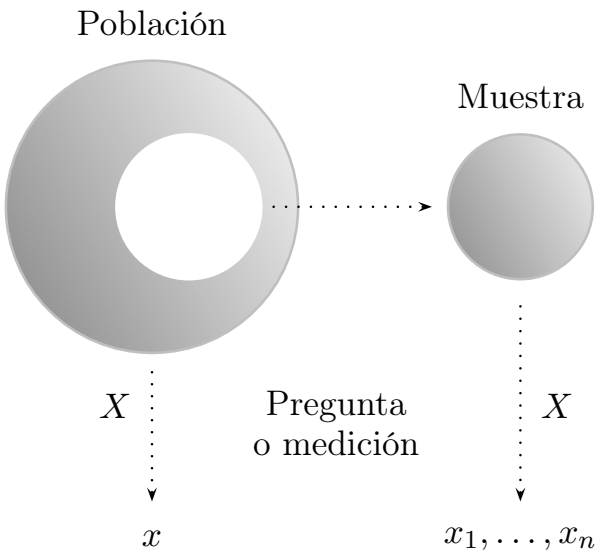


Figura A.1: Una variable aleatoria
y su aplicación a los elementos de una muestra.

Existen varios tipos de variables aleatorias dependiendo del conjunto de valores que estas toman. Dos grandes grupos son las discretas y las continuas. Se dice que una variable aleatoria es discreta si toma un número finito de valores o bien toma un número infinito pero numerable, por ejemplo, $\{0, 1, 2, \ldots\}$. En cambio, una variable aleatoria es continua si puede tomar todos los valores dentro de un cierto intervalo de número reales, por ejemplo, el intervalo $(0, \infty)$. Esta última definición no es del todo correcta pero no es conveniente ahora profundizar en detalles técnicos.

Cada variable aleatoria tiene asociada una distribución de probabilidad (o simplemente una distribución) que establece los valores que puede tomar la

variable aleatoria y las probabilidades con las que toma estos valores. Por ejemplo, si X es discreta y toma los valores $x_1, x_2, \ldots$, entonces su distribución está dada por la especificación de las probabilidades $P(X = x)$ para $x = x_1, x_2, \ldots$ A la función definida por $f(x) = P(X = x)$ se le llama función de probabilidad de la variable aleatoria o de la distribución. Por otro lado, cuando X es continua, su distribución se especifica por las probabilidades $P(X \in (a, b))$, para cualquier intervalo (a, b) de números reales. Cuando estas probabilidades se pueden expresar como una integral sobre una cierta función $f(x)$ sobre el intervalo (a, b), a la función $f(x)$ se le llama función de densidad de la variable aleatoria o de la distribución.

Existen varias distribuciones de importancia que tienen nombre propio. Dentro de las distribuciones discretas podemos mencionar las siguientes: Bernoulli, binomial, Poisson, geométrica, etc. Por ejemplo, la distribución Bernoulli toma únicamente dos valores, 0 y 1, y su distribución se especifica de la forma siguiente:

$$\begin{aligned} P(X = 0) &= 1 - p, \\ P(X = 1) &= p, \end{aligned}$$

en donde p es un parámetro que puede tomar cualquier valor en el intervalo $(0, 1)$. Dentro de las distribuciones continuas se encuentran: exponencial, gamma, normal, entre muchas otras. Posiblemente la distribución normal sea la más importante y la más conocida de todas ellas. Esta distribución se especifica de la forma siguiente: la probabilidad de que la variable aleatoria X tome un valor en un intervalo (a, b) está dada por la siguiente integral

$$P(X \in (a, b)) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\, dx,$$

en donde μ y σ^2 son dos parámetros. En el integrando aparece la función de densidad $f(x)$ de la distribución normal, y cuya gráfica en forma de campana y ampliamente conocida se muestra en la Figura A.2. En esta misma figura se muestra geométricamente que la probabilidad está dada por el área bajo la curva normal en el intervalo (a, b).

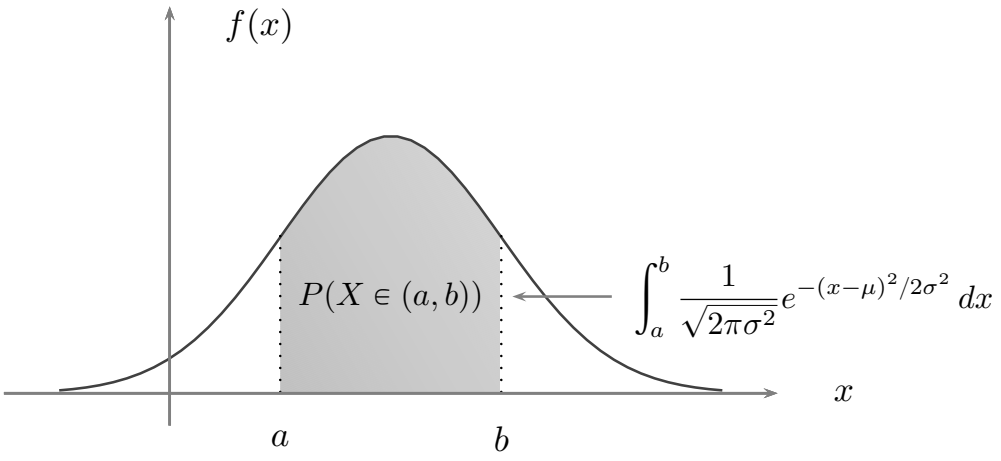


Figura A.2: El área bajo una curva como una probabilidad.

Toda variable aleatoria tiene asociada una función en extremo importante: la función de distribución de una variable aleatoria discreta o continua X se define como la función $F(x) = P(X \leqslant x)$, en donde es necesario notar que la letra F es mayúscula. Se trata de la acumulación de la probabilidad hasta un valor x cualquiera y esta expresión es similar a la que aparece como función de distribución empírica para un conjunto de datos numéricos que hemos mencionado antes. Desde el punto de vista matemático, la función de distribución es importante pues contiene toda la información de la variable aleatoria: sus valores y sus probabilidades.

Con los elementos anteriores podemos extender las definiciones de las descripciones numéricas para conjuntos de datos numéricos a variables aleatorias. Escribiremos las expresiones en el caso continuo y por lo tanto usaremos integrales. En el caso discreto las integrales se reemplazan por sumas y la función de densidad se reemplaza por la función de probabilidad.

La media o esperanza de una variable aleatoria X es un número que se denota por $E(X)$ y corresponde a su valor esperado o valor promedio. De manera breve se usa también la letra μ (se lee mu) para denotarla y se calcula de la siguiente manera:

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

Es muy útil considerar la esperanza de una función de una variable aleatoria, esto es, si $\varphi(X)$ es una función aplicada a la variable aleatoria X, entonces puede demostrarse que su esperanza se calcula como sigue:

$$E(\varphi(X)) = \int_{-\infty}^{\infty} \varphi(x)\, f(x)\, dx.$$

En la tabla que aparece a continuación se encuentran algunas descripciones numéricas para variables aleatorias. Aparecen en el mismo orden en el que estudiamos las cantidades análogas para datos numéricos. Compare estas definiciones con las que aparecen en la tabla de la página 77.

RESUMEN DE FÓRMULAS

Descripciones numéricas para una variable aleatoria X
con función de densidad o de probabilidad $f(x)$

Media	$\mu = E(X) = \displaystyle\int_{-\infty}^{\infty} x\, f(x)\, dx$
Moda	Valor x en donde $f(x)$ es máxima
Mediana	Valor m tal que $P(X \leqslant m) \geqslant 1/2$ y $P(X \geqslant m) \geqslant 1/2$
Varianza	$\sigma^2 = E(X-\mu)^2 = \displaystyle\int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx.$
Desviación estándar	$\sigma = \sqrt{E(X-\mu)^2}.$
Desviación media	$E\lvert X-\mu\rvert = \displaystyle\int_{-\infty}^{\infty} \lvert x-\mu\rvert\, f(x)\, dx.$
Rango	Conjunto de valores de la v.a.
Coeficiente de variación	$\sigma/\mu.$
Momentos	$\mu'_k = E(X^k) = \displaystyle\int_{-\infty}^{\infty} x^k\, f(x)\, dx.$
Momentos centrales	$\mu_k = E(X-\mu)^k = \displaystyle\int_{-\infty}^{\infty} (x-\mu)^k\, f(x)\, dx.$
Cuantil al $100p\,\%$	Valor x tal que $P(X \leqslant x) \geqslant p$ y $P(X \geqslant x) \geqslant 1-p$
Asimetría	μ_3/σ^3
Curtosis	μ_4/σ^4

Sugerencias a los ejercicios

Esta sección contiene algunas sugerencias de solución a los ejercicios planteados. Para algunos ejercicios es necesario ser más explícito al dar una solución o al justificar una respuesta, considere por tanto que este material contiene simplemente ideas para generar una solución completa, correcta y bien escrita. La mayoría de las gráficas han sido omitidas. Recuerde además que los métodos empleados o sugeridos para llegar a una solución no son necesariamente únicos.

1. Se puede tomar como población

   a) el conjunto de votantes. Una unidad de observación puede ser los resultados de las casillas electorales.

   b) el conjunto de mediciones de contaminantes en el aire. Una unidad de observación puede ser una medición de un cierto contaminante en una cierta zona geográfica de la ciudad.

   c) el conjunto de todas las personas fumadoras. Una unidad de observación puede ser una persona fumadora.

   d) el conjunto de usuarios de la biblioteca, o bien el conjunto de materiales de la biblioteca. Una unidad de observación puede ser un usuario de la biblioteca, o bien un material de la biblioteca.

   e) el conjunto de todas las familias que habitan en la región geográfica. Una unidad de observación puede ser una familia.

   f) la totalidad de los productos de la fábrica. Una unidad de observación puede ser un producto o un lote de productos.

   g) el conjunto de personas enfermas. Una unidad de observación puede ser una persona enferma.

   h) el conjunto de personas adictas a alguna droga. Una unidad de observación puede ser una persona adicta.

i) el conjunto de casas con servicio de agua. Una unidad de observación puede ser una casa o una colección de ellas.

j) el conjunto de personas que viven en ese país. Una unidad de observación puede ser una persona o una agrupación de ellas.

k) el conjunto de personas en la población en estudio. Una unidad de observación puede ser una persona.

l) el conjunto de personas que ven televisión. Una unidad de observación puede ser una persona con esta característica, o bien un hogar con televisión.

m) el conjunto de niños en la población en estudio. Una unidad de observación puede ser un niño en este conjunto.

n) el conjunto de personas que viven en esa ciudad. Una unidad de observación puede ser una persona en este conjunto.

ñ) el conjunto de personas de edad avanzada. Una unidad de observación puede ser una persona en este conjunto.

2. *a*) La población puede ser el conjunto de personas que viven en ese país. Una unidad de observación puede ser una de estas personas. Tres variables que podrían ser de interés son: edad, estatura y peso de cada persona.

b) La población puede ser el conjunto de habitantes de la ciudad. Una unidad de observación puede ser la población completa y tres variables que podrían ser de interés son: número de robos en el transporte público en un periodo determinado, número de agresiones físicas y verbales, número de robos a comercios o a casas habitación.

c) La población puede considerarse como el conjunto de estudiantes. Una unidad de observación puede ser un estudiante cualquiera. Tres variables que podrían ser de interés son: promedio, porcentaje de avance, número de asignaturas reprobadas.

d) La población puede considerarse como el conjunto de personas en edad avanzada. Una unidad de observación puede ser cualquier persona dentro de esta categoría. Tres variables que podrían ser de interés son: edad, ingresos económicos, saldo en su cuenta bancaria si la tuviera.

e) La población puede ser considerada como la totalidad de personas extranjeras viviendo en el país de referencia. Una unidad de observación puede ser cualquiera de estas personas. Tres variables que podrían ser de interés son: nacionalidad, tiempo transcurrido desde su último ingreso al país, y tipo de visa o permiso de estancia en el país.

f) La población puede ser considerada como el conjunto de habitantes de la ciudad en cuestión. Una unidad de observación puede ser cualquiera de estas personas. Tres variables que podrían ser de interés son: edad, tipos de transporte público utilizados, tiempos de traslado.

g) La población puede ser el conjunto de niños en una ciudad o en una región geográfica. Una unidad de observación puede ser un niño cualquiera. Tres variables que podrían ser de interés son: edad, tiempo promedio frente al televisor durante el día, y cantidad promedio de alimentos poco saludables consumidos en un día.

h) La población es el conjunto de niños recién nacidos. Uno de estos niños puede ser considerado una unidad de observación. Tres variables que podrían ser de interés son: sexo, estatura y peso.

3. *a*) Estado de un artículo.

b) Fruta preferida por una persona.

c) Código postal del domicilio de una persona.

d) Ocupación de una persona económicamente activa.

e) Promedio escolar de un estudiante.

f) Número de infracciones de un conductor.

g) Primera lengua natal de una persona.

h) Edad de una persona.

i) Número de veces que una persona ha contraído matrimonio.

j) Número de automóviles que posee una persona.

k) Primer apellido de la madre de una persona.

l) Hora de la mañana en la que una persona se despierta.

m) Estación del año preferida por una persona.

n) Nivel de inseguridad en una zona de la ciudad.

ñ) Número de horas promedio de sueño de una persona.

o) Tiempo promedio de traslado de una persona para ir de un lugar a otro en una ciudad.

p) Color del coche de una persona con únicamente un coche.

q) Número de días necesarios para concluir un trámite administrativo.

4. Las respuestas no son necesariamente únicas. En cada caso se pueden dar consideraciones para justificar una respuesta dada.

a) Cuantitativa discreta.
b) Cuantitativa continua.
c) Cuantitativa discreta.
d) Cuantitativa continua.
e) Cuantitativa discreta.
f) Cualitativa.
g) Cuantitativa continua.
h) Cuantitativa continua.
i) Cualitativa.
j) Cuantitativa continua.
k) Cuantitativa continua.
l) Cuantitativa discreta.
m) Cualitativa.
n) Cualitativa.
ñ) Cuantitativa discreta.
o) Cuantitativa discreta.
p) Cualitativa.
q) Cualitativa.
r) Cuantitativa discreta.
s) Cuantitativa discreta.
t) Cuantitativa discreta.
u) Cualitativa.
v) Cuantitativa continua.

5. En algunos casos puede se puede considerar un tipo de escala o el otro.

a) Nominal.
b) Ordinal.
c) Ordinal.
d) Ordinal.
e) Nominal.
f) Nominal.
g) Nominal.
h) Ordinal.
i) Ordinal.
j) Nominal.
k) Ordinal.
l) Nominal.
m) Ordinal.
n) Nominal.
ñ) Nominal
o) Nominal.
p) Nominal.
q) Ordinal.

6. Recuerde que las respuestas no son necesariamente únicas y que, bajo algunas consideraciones, una variable puede clasificarse de un tipo o de otro.

a) Discreta, escala de razón.
b) Discreta, escala de razón.
c) Discreta, escala de razón.
d) Discreta, escala de razón.
e) Discreta, escala de razón.
f) Continua, escala de razón.
g) Discreta, escala de razón.
h) Discreta, escala de razón.
i) Discreta, escala de razón.
j) Discreta, escala de razón.
k) Discreta, escala de razón.
l) Discreta, escala de razón.
m) Discreta, escala de razón.
n) Discreta, escala de razón.
ñ) Discreta, escala de razón.

7. Suponga la siguiente clasificación

$$A = [0, 13), B = [13, 20), C = [20, 35), D = [35, 65), E = [65, \infty).$$

   Se trata de una variable cualitativa con escala de medición ordinal. Tomando aproximadamente el punto medio de cada intervalo, se pueden definir los siguientes representantes de clase: 6.5, 16.5, 27.5, 50 y 82.5, respectivamente.

8. Se trata de una variable cualitativa con escala de medición ordinal. Tomando aproximadamente el punto medio de cada intervalo, se pueden definir los siguientes representantes de clase: 2.5, 7.5 y 20, respectivamente.

9. Se trata de una variable cualitativa con escala de medición nominal. Tomando el punto medio de cada conjunto de valores, se pueden definir los siguientes representantes de clase: 6 para la categoría par y 5 para la categoría impar.

10. Se trata de una variable cualitativa con escala de medición nominal. Considerando una totalidad de 5 continentes y tomando como país representante el país más poblado por continente, los representantes de clase son: China (Asia), Estados Unidos (América), Nigeria (África), Rusia (Europa) y Australia (Oceanía).

11. Los siguientes valores fueron obtenido en R mediante la función `mean()`.

    a) 1.625 .
    b) 3.045455 .
    c) 27 .
    d) 0.3333333 .

12. Sólo es necesario hacer el primer inciso.

    a) 18 .
    b) 18 .
    c) 18 .
    d) 18 .

13. $x_5 = 8$.

14. Es suficiente hacer el cálculo del conjunto de datos original.

    a) $\bar{x} = 4$.
    b) $\overline{x + 2} = \bar{x} + 2 = 6$.
    c) $\overline{x - 4} = \bar{x} - 4 = 0$.
    d) $\overline{2x} = 2 \cdot \bar{x} = 8$.
    e) $\overline{10x} = 10 \cdot \bar{x} = 40$.

15. $\bar{x} = -0.3529412$.

16. $\bar{x} = 8.692308$.

17. $\bar{x} = 2.5$ en el primer caso y $\bar{x} = 0$ en el segundo caso. No es necesario hacer mayores cálculos.

18. $\bar{x} = 4.322581$ y $\bar{x} = 0.5833333$.

19. No es necesario hacer cálculos.

    a) Ese mismo dato.
    b) Ese mismo dato.
    c) El promedio de ellos.
    d) Ese mismo dato.

20. *a*) Verdadero.
    b) Verdadero.
    c) Falso.
    d) Puede ser falso o verdadero dependiendo del dato añadido. Si el dato adicional es menor a la media original, la nueva media disminuye. Si el dato adicional es mayor a la media original, la nueva media aumenta. La nueva media no cambia si el dato adicional es la media original.
    e) Misma situación que en el inciso anterior.

21. El análisis es similar al presentado en el texto.

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\,(y_1 + \cdots + y_n) \\
&= \frac{1}{n}\,((ax_1 + c) + \cdots + (ax_n + c)) \\
&= \frac{1}{n}\,[\,a(x_1 + \cdots + x_n) + nc\,] \\
&= a\,\frac{1}{n}\,(x_1 + \cdots + x_n) + c \\
&= a\bar{x} + c.
\end{aligned}$$

22.

$$\begin{aligned}
\bar{x}_{n+1} &= \frac{1}{n+1}\,(x_1 + \cdots + x_{n+1}) \\
&= \frac{1}{n+1}\,(x_1 + \cdots + x_n) + \frac{1}{n+1}\,x_{n+1} \\
&= \frac{n}{n+1}\,\frac{1}{n}\,(x_1 + \cdots + x_n) + \frac{1}{n+1}\,x_{n+1} \\
&= \frac{1}{n+1}\,n\,\bar{x} + \frac{1}{n+1}\,x_{n+1} \\
&= \frac{1}{n+1}\,(n\,\bar{x}_n + x_{n+1}).
\end{aligned}$$

23. La misma media. Use la fórmula del ejercicio anterior para comprobar esta afirmación.

24.

$$\begin{aligned}
\bar{x}_{n-1} &= \frac{1}{n-1}\,(x_1 + \cdots + x_{i-1} + x_{i+1} + \cdots + x_n) \\
&= \frac{1}{n-1}\,(x_1 + \cdots + x_n - x_i) \\
&= \frac{1}{n-1}\,(x_1 + \cdots + x_n) - \frac{1}{n-1}\,x_i \\
&= \frac{n}{n-1}\,\frac{1}{n}\,(x_1 + \cdots + x_n) - \frac{1}{n-1}\,x_i \\
&= \frac{1}{n-1}\,n\,\bar{x} - \frac{1}{n-1}\,x_i \\
&= \frac{1}{n-1}\,(n\,\bar{x}_n - x_i).
\end{aligned}$$

25.

$$\begin{aligned}
\frac{1}{n+m}\,((x_1 + \cdots + x_n) + (y_1 + \cdots + y_m)) &= \frac{n}{n+m}\,\frac{1}{n}\,(x_1 + \cdots + x_n) \\
&\quad + \frac{m}{n+m}\,\frac{1}{m}\,(y_1 + \cdots + y_m) \\
&= \frac{n}{n+m}\,\bar{x} + \frac{m}{n+m}\,\bar{y}.
\end{aligned}$$

26. Sea $\bar{x} = 8$ la media de calificaciones durante el primer semestre y sea $\bar{y}$ la media desconocida de calificaciones del segundo semestre. Entonces

$$\frac{5}{5+4}\,\bar{x} + \frac{4}{5+4}\,\bar{y} = 8.5\,.$$

Substituyendo el valor de $\bar{x}$ se obtiene $\bar{y} = 9.125\,$.

27. Se usan las propiedades de la sumatoria.

a) $\displaystyle\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$

b) $\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2$

$\displaystyle= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$

28. Falso.

29. $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \dfrac{1}{n}\sum_{i=1}^{n} \dfrac{x_i}{\bar{x}} = \dfrac{1}{\bar{x}} \cdot \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{\bar{x}} \cdot \bar{x} = 1.$

30. *a*) Aplique las leyes del logaritmo.

    b) $\text{mg}(ax) = \sqrt[n]{ax_1 \cdots ax_n} = \sqrt[n]{a^n(x_1 \cdots x_n)} = (a^n)^{1/n}\sqrt[n]{x_1 \cdots x_n}$ $= a \cdot \text{mg}(x)$.

    c) $\text{mg}(x/y) = \sqrt[n]{\dfrac{x_1 \cdots x_n}{y_1 \cdots y_n}} = \dfrac{\sqrt[n]{x_1 \cdots x_n}}{\sqrt[n]{y_1 \cdots y_n}} = \dfrac{\text{mg}(x)}{\text{mg}(y)}$.

31. *a*) Evidente.

    b) Simplemente multiplique y divida $h(x)$ por $(x_1 \cdots x_n)$.

32. *a*) El conjunto de modas es $\{a, c\}$.

    b) El conjunto de modas es $\{a, b, c\}$.

    c) El conjunto de modas es $\{b, d\}$.

    d) No hay moda.

33. Se usa el símbolo Moda(x) para denotar a la posible colección de modas del conjunto de datos x.

    a) $\text{Moda}(x) = \{3\}$.

    b) $\text{Moda}(x + 3) = \text{Moda}(x) + 3 = \{6\}$.

    c) $\text{Moda}(x/2) = (1/2) \cdot \text{Moda}(x) = \{3/2\}$.

    d) $\text{Moda}(x - 2) = \text{Moda}(x) - 2 = \{1\}$.

    e) $\text{Moda}(2x) = 2 \cdot \text{Moda}(x) = \{6\}$.

    f) $\text{Moda}(10x) = 10 \cdot \text{Moda}(x) = \{30\}$.

34. Falso.

35. Los siguientes valores fueron obtenido en R mediante la función `median()`.

    a) $\tilde{x} = 1.5$.

    b) $\tilde{x} = 2.7$.

    c) $\tilde{x} = 28$.

    d) $\tilde{x} = 0$.

36. Es suficiente con hacer el primer inciso.

a) $\tilde{x} = 15$.
b) $\tilde{x} = 15$.
c) $\tilde{x} = 15$.
d) $\tilde{x} = 15$.

37. Se usa el símbolo med(x) para denotar a la mediana del conjunto de datos x.
$\text{med}(x) = 3.5$.
$\text{med}(x + 2) = \text{med}(x) + 2 = 5.5$.
$\text{med}(x + 4) = \text{med}(x) + 4 = 7.5$.
$\text{med}(x - 3) = \text{med}(x) - 3 = 0.5$.
$\text{med}(x/2) = (1/2)\,\text{med}(x) = 1.75$.
$\text{med}(5x) = 5 \cdot \text{med}(x) = 17.5$.

38. No es necesario hacer ningún cálculo.

a) Ese mismo dato.
b) Ese mismo dato.
c) El promedio de esos datos.
d) Ese mismo dato.
e) El dato de en medio.
f) Ese mismo dato.

39. La misma mediana.

40. *a*) Verdadero.
b) verdadero.

41. *a*) $\tilde{x} = 2$.
b) $\tilde{x} = 1$.

42. *a*) La mediana es el número correspondiente al tercer punto de izquierda a derecha.
b) La mediana es el número correspondiente al segundo punto de izquierda a derecha.

43. Falso.

44. *a*) La media aumenta.
b) La moda pasa de no existir a existir, y toma el valor 8.
c) La mediana cambia de 4 a 5.

45. Los siguientes resultados fueron obtenidos en el paquete R, y por lo tanto, el promedio no es sobre el número de datos n, sino sobre $n - 1$.

a) $\text{var}(x) = 1$.

b) $\text{var}(x) = 0$.

c) $\text{var}(x) = 1$.

d) $\text{var}(x) = 1$.

46. Los valores son -1 y 1.

47. *a*) Por definición,

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\,\frac{1}{n}\sum_{i=1}^{n}x_i + \bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}^2 + \bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2. \end{aligned}$$

b) Por definición,

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[(x_i - c) - (\bar{x} - c)]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[(x_i - c)^2 - 2(x_i - c)(\bar{x} - c) + (\bar{x} - c)^2] \\ &= [\,\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2\,] - 2(\bar{x} - c)(\bar{x} - c) + (\bar{x} - c)^2 \\ &= [\,\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2\,] - (\bar{x} - c)^2. \end{aligned}$$

48. Recordemos que $\bar{y} = a\bar{x} + c$. Esto es usado en el siguiente cálculo.

$$\begin{aligned} s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}((ax_i + c) - (a\bar{x} + c))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}a^2 \cdot (x_i - \bar{x})^2 \\ &= a^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= a^2 \cdot s_x^2. \end{aligned}$$

49. Es suficiente hacer el primer inciso. El cálculo fue realizado en R y, por lo tanto, el promedio no es sobre el número de datos $n = 5$, sino sobre $n-1 = 4$.

    a) $\text{var}(x) = 2.8$.

    b) $\text{var}(x + 1) = 2.8$.

    c) $\text{var}(x - 2) = 2.8$.

    d) $\text{var}(2x) = 4 \cdot \text{var}(x) = 11.2$.

    e) $\text{var}(x/2) = (1/4) \cdot \text{var}(x) = 0.7$.

    f) $\text{var}(2x + 1) = \text{var}(2x) = 4 \cdot \text{var}(x) = 11.2$.

50. La solución de este ejercicio puede requerir saber un poco de cálculo diferencial. Usando derivadas puede comprobarse que la función cuadrática $g(u)$ alcanza un valor mínimo en $u = \bar{x}$. Compruebe que $g'(u) = 0 \Leftrightarrow u = \bar{x}$, y que $g''(u) = 2n > 0$.

51. *a*) Verdadero.

    b) Verdadero.

    c) Falso.

52. *a*) $\bar{y} = 8, \quad s_y^2 = 48.$

    b) $\bar{y} = -3, \quad s_y^2 = 12.$

    c) $\bar{y} = 1, \quad s_y^2 = 12.$

    d) $\bar{y} = -8/3, \quad s_y^2 = 16/9.$

53. Puede comprobarse con facilidad que $\bar{y} = 1$. Por lo tanto,

$$\begin{aligned} s_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i}{\bar{x}} - 1)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{\bar{x}})^2 \\ &= \frac{1}{\bar{x}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{\bar{x}^2}\cdot s_x^2. \end{aligned}$$

Alternativamente, observe que $\bar{x}$ es una constante respecto del subíndice i y, por lo tanto, se separa de la varianza al cuadrado.

54. La media es $(1.10)\cdot(\$550.00) = \$605.00$, y la varianza es $(1.10)^2\cdot(\$50)^2 = (\$55)^2$.

55. Esta propiedad se sigue directamente de la propiedad correspondiente para la varianza. Tenemos que

$$s_y^2 = a^2\cdot s_x^2.$$

Tomando raíz cuadrada y observando que $\sqrt{a^2} = |a|$ se obtiene la igualdad buscada.

56. Los siguientes cálculos fueron realizados en R y, por lo tanto, el promedio no se efectúa sobre el número de datos $n = 4$, sino sobre $n - 1 = 3$.

a) $\mathrm{sd}(x) = 0.8164966$.
b) $\mathrm{sd}(|x|) = 0.5773503$.
c) $\mathrm{sd}(x + 2) = \mathrm{sd}(x) = 0.8164966$.
d) $\mathrm{sd}(3x - 2) = \mathrm{sd}(3x) = 3\cdot\mathrm{sd}(x) = 2.44949$.

57. Para la media tenemos que

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})/s_x = [(\frac{1}{n}\sum_{i=1}^{n}x_i) - \bar{x}]/s_x = [\bar{x} - \bar{x}]/s_x = 0.$$

Entonces la varianza es

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2/s_x^2 = s_x^2/s_x^2 = 1.$$

58. *a*) $\mathrm{dm}(x) = 1$.

b) $\mathrm{dm}(x+2) = \mathrm{dm}(x) = 1$.

c) $\mathrm{dm}(x-3) = \mathrm{dm}(x) = 1$.

d) $\mathrm{dm}(2x) = 2 \cdot \mathrm{dm}(x) = 2$.

e) $\mathrm{dm}(x/2) = (1/2) \cdot \mathrm{dm}(x) = 1/2$.

f) $\mathrm{dm}(-5x) = |-5| \cdot \mathrm{dm}(x) = 5$.

59. Sea $y = ax + c$. Recordemos que $\bar{y} = a\bar{x} + c$. Esto es usado en el siguiente cálculo.

$$\begin{aligned}
\mathrm{dm}(y) &= \frac{1}{n}\sum_{i=1}^{n} |y_i - \bar{y}| \\
&= \frac{1}{n}\sum_{i=1}^{n} |(ax_i + c) - (a\bar{x} + c)| \\
&= \frac{1}{n}\sum_{i=1}^{n} |ax_i - a\bar{x}| \\
&= \frac{1}{n}\sum_{i=1}^{n} |a| \cdot |x_i - \bar{x}| \\
&= |a| \cdot \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \\
&= |a| \cdot \mathrm{dm}(x).
\end{aligned}$$

60. Es suficiente hacer el primer cálculo.

a) $r(x) = 10$.

b) $r(x) = 10$.

c) $r(x) = 10$.

d) $r(x) = 10$.

61. *a*) Verdadero.

b) Falso.

c) Verdadero.

d) Verdadero.

62. Denotemos por x^* al valor más grande dentro del conjunto $\{x_1, \ldots, x_n\}$, y sea x_* el valor más pequeño. La comprobación de la fórmula se divide en tres casos. Supongamos primero que $a > 0$. Entonces el valor más grande de

$ax_1 + c, \ldots, ax_n + c$ es $ax^* + c$ y el valor más pequeño es $ax_* + c$. Por lo tanto,

$$\begin{aligned} \text{Rango}(ax + c) &= (ax^* + c) - (ax_* + c) \\ &= ax^* - ax_* \\ &= a \cdot (x^* - x_*) \\ &= |a| \cdot \text{Rango}(x). \end{aligned}$$

Si $a = 0$, entonces el conjunto de datos transformados es $c, \ldots, c$, cuyo rango es

$$\text{Rango}(ax + c) = c - c = 0 = |a| \cdot \text{Rango}(x).$$

Finalmente supongamos que $a < 0$. Entonces el valor más grande de $ax_1 + c, \ldots, ax_n + c$ es $ax_* + c$ y el valor más pequeño es $ax^* + c$. Estas observaciones son aquí cruciales. Por lo tanto,

$$\begin{aligned} \text{Rango}(ax + c) &= (ax_* + c) - (ax^* + c) \\ &= ax_* - ax^* \\ &= -a \cdot (x^* - x_*) \\ &= |a| \cdot \text{Rango}(x). \end{aligned}$$

63. *a*) Sí puede serlo. Lo es cuando $s > 0$ y $\bar{x} < 0$.

b) Sí puede serlo. Lo es cuando $s = 0$ y $\bar{x} \neq 0$.

c) Que se trata de un conjunto de datos constante no cero.

64. Puede comprobarse con facilidad que $\bar{y} = 1$. Por lo tanto,

$$\begin{aligned}
\mathrm{cv}(y) &= s_y/\bar{y} \\
&= s_y \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{x_i}{\bar{x}} - 1)^2} \\
&= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{\bar{x}})^2} \\
&= \sqrt{\frac{1}{\bar{x}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \sqrt{\frac{1}{\bar{x}^2}}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{1}{|\bar{x}|}\cdot s_x\,.
\end{aligned}$$

65. *a)* $m_1 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^1 = \frac{1}{n}\sum_{i=1}^{n} x_i - \bar{x} = 0.$

b) $m_2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2.$

c) Este cálculo ha sido desarrollado antes en varias ocasiones.

$$\begin{aligned} m_2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}\,\frac{1}{n}\sum_{i=1}^{n}x_i+\bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2-2\bar{x}^2+\bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2 \\ &= m_2'-(m_1')^2. \end{aligned}$$

66. Recordemos que la media del conjunto de datos trasladados $x+c$ es $\bar{x}+c$. Entonces

$$m_k(x+c)=\frac{1}{n}\sum_{i=1}^{n}((x_i+c)-(\bar{x}+c))^k=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k=m_k(x).$$

67. Recordemos que la media del conjunto de datos modificados ax es $a\bar{x}$. Entonces

$$m_k(ax)=\frac{1}{n}\sum_{i=1}^{n}(ax_i-a\bar{x})^k=\frac{1}{n}\sum_{i=1}^{n}a^k\cdot(x_i-\bar{x})^k=a^k\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k=a^k\cdot m_k(x).$$

68. Lleve a cabo los cocientes indicados.

Valor	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
A	2	2	2/28	2/28
B	8	10	8/28	10/28
C	6	16	6/28	16/28
D	4	20	4/28	20/28
E	3	23	3/28	23/28
F	5	28	5/28	28/28

69. Se inicia colocando la frecuencia relativa faltante, que es 0.3, en el tercer renglón. Sumando se puede completar la cuarta columna. Considerando que son 20 observaciones, se completa la primera columna multiplicando 20 por la frecuencia relativa. Después se calcula la segunda columna.

Valor	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
A	2	2	0.1	0.1
B	8	10	0.4	0.5
C	6	16	0.3	0.8
D	4	20	0.2	1

70. Sea p un número en el intervalo $(0, 1]$. Un cuantil es una cantidad que separa un conjunto de datos numéricos ordenados en dos partes dejando a la izquierda una proporción por lo menos del $100p\,\%$ de los datos, y la proporción restante $100(1-p)\,\%$, por lo menos, de los datos quedan a la derecha. El caso particular cuando

a) $p = 0.25, 0.5$ ó 0.75, al cuantil se le llama (primer, segundo, tercer) cuartil.

b) $p = 0.01, 0.02, \ldots, 0.99$, al cuantil se le llama (primer, segundo,...) percentil.

71. Los siguientes resultados fueron obtenidos en R.

a) $c_{0.25} = -1, \quad c_{0.5} = 0, \quad c_{0.75} = 1\,.$

b) $c_{0.25} = 2, \quad c_{0.5} = 2, \quad c_{0.75} = 2\,.$

c) $c_{0.25} = 0, \quad c_{0.5} = 1.5, \quad c_{0.75} = 3\,.$

d) $c_{0.25} = 2.3, \quad c_{0.5} = 2.7, \quad c_{0.75} = 3.7\,.$

e) $c_{0.25} = 25, \quad c_{0.5} = 28, \quad c_{0.75} = 30\,.$

72. Los siguientes resultados fueron obtenidos en R.

a) $c_{0.2} = 3, \quad c_{0.4} = 3, \quad c_{0.6} = 5, \quad c_{0.8} = 5\,.$

b) $c_{0.2} = 2, \quad c_{0.4} = 4, \quad c_{0.6} = 4, \quad c_{0.8} = 6\,.$

73. Recordemos que la media de los datos trasladados $x+c$ es $\bar{x}+c$. Por lo tanto,

$$\begin{aligned}
\mathrm{sk}(x+c) &= \frac{1}{s^3(x+c)} \left(\frac{1}{n} \sum_{i=1}^{n} ((x_i + c) - (\bar{x} + c))^3 \right) \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} ((x_i + c) - (\bar{x} + c))^3}{(\frac{1}{n}\sum_{i=1}^{n} ((x_i + c) - (\bar{x} + c))^2)^{3/2}} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2)^{3/2}} \\
&= \mathrm{sk}(x).
\end{aligned}$$

74. Recordemos que la media de los datos ax es $a\bar{x}$. Por lo tanto,

$$\begin{aligned}
\text{sk}(ax) &= \frac{1}{s^3(ax)}\left(\frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^3\right) \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^2)^{3/2}} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(a(x_i - \bar{x}))^3}{(\frac{1}{n}\sum_{i=1}^{n}(a(x_i - \bar{x}))^2)^{3/2}} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}a^3(x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}a^2\cdot(x_i - \bar{x})^2)^{3/2}} \\
&= \frac{a^3\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(a^2)^{3/2}\cdot(\frac{1}{n}\sum_{i=1}^{n}\cdot(x_i - \bar{x})^2)^{3/2}} \\
&= \frac{a}{|a|}\cdot\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^{3/2}} \\
&= \begin{cases} \text{sk}(x) & \text{si } \ a > 0, \\ -\text{sk}(x) & \text{si } \ a < 0. \end{cases}
\end{aligned}$$

75. Los siguientes cálculos fueron hechos en R.

   a) $k(x) = 1.5$.

   b) $k(x) = 1.5$.

   c) $k(x) = 1.5$.

76. Recordemos que la media del conjunto de datos $x + c$ es $\bar{x} + c$. Entonces, por definición,

$$\begin{aligned}
k(x+c) &= \frac{1}{s^4(x+c)}\left(\frac{1}{n}\sum_{i=1}^{n}((x_i + c) - (\bar{x} + c))^4\right) \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}((x_i + c) - (\bar{x} + c))^4}{(\frac{1}{n}\sum_{i=1}^{n}((x_i + c) - (\bar{x} + c))^2)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \\
&= k(x).
\end{aligned}$$

77. Recordemos que la media del conjunto de datos ax es $a\bar{x}$. Entonces, por

definición,

$$\begin{aligned} k(ax) &= \frac{1}{s^4(ax)}\left(\frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^4\right) \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(ax_i - a\bar{x})^2)^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} a^4 \cdot (x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n} a^2 \cdot (x_i - \bar{x})^2)^2} \\ &= \frac{a^4 \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{a^4 \cdot (\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} \\ &= k(x). \end{aligned}$$

78. Supongamos que tomamos como las marcas de clase los valores $0, 1.5, 3.5$ y 5.5, respectivamente. Sus frecuencias son $3, 5, 2$ y 4. Entonces la media, moda y mediana son:

    a) $\bar{x} = 2.607143$.

    b) $\text{Moda}(x) = 1.5$.

    c) $\tilde{x} = 1.5$.

79. Consulta de una publicación.

80. Gráfica omitida.

81. Gráfica omitida.

82. Gráfica omitida.

83. Gráfica omitida.

84. Investigación y gráfica omitidas.

85. Investigación y gráfica omitidas.

86. Investigación y gráfica omitidas.

87. En ambos tipos de gráficas los datos pueden ser cualitativos o cuantitativos. Sin embargo, en un histograma existe un orden entre los valores de la variable.

88. Gráfica omitida.

89. Gráfica omitida.

90. Gráfica omitida.

91. Gráfica omitida.

92. Gráfica omitida.

93. Gráfica omitida.

94. Gráfica omitida.

95. Gráfica omitida.

96. Gráfica omitida.

97. Gráfica omitida.

98. Gráfica omitida.

99. Gráfica omitida.

100. Sólo se proporciona la expresión analítica de cada función de distribución empírica. Se omiten las gráficas correspondientes.

$a)\ F(x) = \begin{cases} 0 & \text{si } x < 2, \\ 1/2 & \text{si } 2 \leqslant x < 5, \\ 1 & \text{si } x \geqslant 5. \end{cases}$

$b)\ F(x) = \begin{cases} 0 & \text{si } x < -1, \\ 1/3 & \text{si } -1 \leqslant x < 0, \\ 2/3 & \text{si } 0 \leqslant x < 1, \\ 1 & \text{si } x \geqslant 1. \end{cases}$

$c)\ F(x) = \begin{cases} 0 & \text{si } x < 0, \\ 2/6 & \text{si } 0 \leqslant x < 1, \\ 3/6 & \text{si } 1 \leqslant x < 2, \\ 4/6 & \text{si } 2 \leqslant x < 3, \\ 5/6 & \text{si } 3 \leqslant x < 5, \\ 1 & \text{si } x \geqslant 5. \end{cases}$

$d)\ F(x) = \begin{cases} 0 & \text{si } x < 4, \\ 3/6 & \text{si } 4 \leqslant x < 10, \\ 1 & \text{si } x \geqslant 10. \end{cases}$

e) $F(x) = \begin{cases} 0 & \text{si } x < 7, \\ 1 & \text{si } x \geqslant 7. \end{cases}$

f) $F(x) = \begin{cases} 0 & \text{si } x < 25, \\ 1 & \text{si } x \geqslant 25. \end{cases}$

101. *a*) El conjunto de datos es $1, 1, 2, 3$. Cualquier conjunto de datos con la misma proporción de estos valores produce la misma función de distribución empírica. Esta función es

$$F(x) = \begin{cases} 0 & \text{si } x < 1, \\ 2/4 & \text{si } 1 \leqslant x < 2, \\ 3/4 & \text{si } 2 \leqslant x < 3, \\ 1 & \text{si } x \geqslant 3. \end{cases}$$

b) El conjunto de datos es $-2, -1, 1, 2$. Cualquier conjunto de datos con la misma proporción de estos valores produce la misma función de distribución empírica. Esta función es

$$F(x) = \begin{cases} 0 & \text{si } x < -2, \\ 1/4 & \text{si } -2 \leqslant x < -1, \\ 2/4 & \text{si } -1 \leqslant x < 1, \\ 3/4 & \text{si } 1 \leqslant x < 2, \\ 1 & \text{si } x \geqslant 2. \end{cases}$$

c) El conjunto de datos es $0, 0, 0, 1, 2, 2, 3, 3$. Cualquier conjunto de datos con la misma proporción de estos valores produce la misma función de distribución empírica. Esta función es

$$F(x) = \begin{cases} 0 & \text{si } x < 0, \\ 3/8 & \text{si } 0 \leqslant x < 1, \\ 4/8 & \text{si } 1 \leqslant x < 2, \\ 6/8 & \text{si } 2 \leqslant x < 3, \\ 1 & \text{si } x \geqslant 3. \end{cases}$$

102. Los cuantiles son $c_{0.2} = 0$, $c_{0.4} = 1$, $c_{0.6} = 3$, $c_{0.8} = 3$. Véase la Figura A.3.

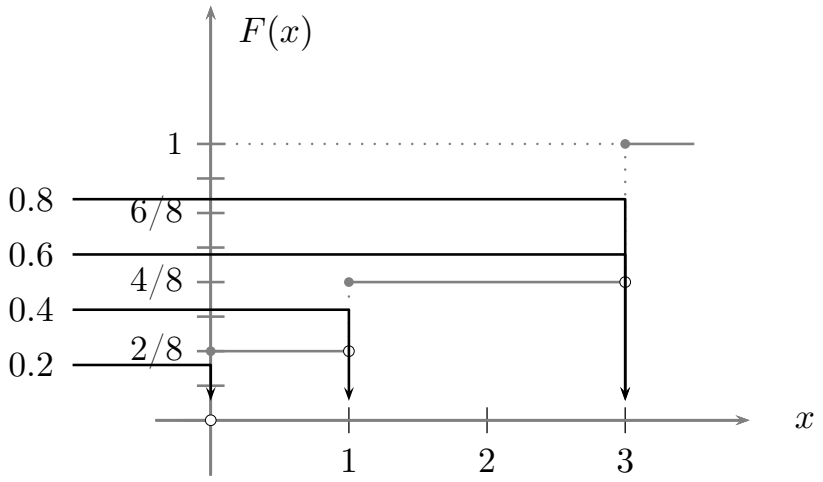


Figura A.3

103. *a*) La correlación es positiva pues cuando x se incrementa, y también parece incrementarse.

b) La correlación es negativa pues cuando x se incrementa, y parece decrementarse.

104. Las frecuencias individuales son

x	0	1	2	3
Frecuencia	16	15	12	9

y	10	20	30	40
Frecuencia	14	9	13	16

105. *a*) $\text{cov}(x, y) = 0.3333$. La función `cov()` en R produce el valor 0.5.

b) $\text{cov}(x, y) = -1.625$. La función `cov()` en R produce el valor -2.166667.

c) $\text{cov}(x, y) = 2.08$. La función `cov()` en R produce el valor 2.6.

106. Recordemos que la media del conjunto de datos $ax + b$ es $a\bar{x} + b$, y para el conjunto $cy + d$ es $c\bar{y} + d$. Entonces

$$\begin{aligned}
\operatorname{cov}(ax+b, cy+d) &= \frac{1}{n}\sum_{i=1}^{n}((ax_i+b)-(a\bar{x}+b))((cy_i+d)-(c\bar{y}+d)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(a(x_i-\bar{x}))(c(y_i-\bar{y})) \\
&= ac\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
&= ac\cdot\operatorname{cov}(x,y).
\end{aligned}$$

107. Por definición,

$$\operatorname{cov}(x,c) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(c-c) = 0.$$

108. *a*) La covarianza es pequeña pues los factores $(x_i - \bar{x})$ son pequeños.

b) La covarianza es pequeña pues los factores $(y_i - \bar{y})$ son pequeños.

109. Este resultado es una consecuencia de las fórmulas $\operatorname{cov}(ax + b, cy + d) = ac \cdot \operatorname{cov}(x, y)$ y $\operatorname{var}(ax + b) = a^2 \cdot \operatorname{var}(x)$. En efecto,

$$\begin{aligned}
\rho(ax+b, cy+d) &= \frac{\operatorname{cov}(ax+b, cy+d)}{\sqrt{\operatorname{var}(ax+b)\cdot\operatorname{var}(cy+d)}} \\
&= \frac{ac\cdot\operatorname{cov}(x,y)}{\sqrt{a^2\cdot\operatorname{var}(x)\cdot c^2\cdot\operatorname{var}(y)}} \\
&= \frac{a}{|a|}\frac{c}{|c|}\cdot\frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)\cdot\operatorname{var}(y)}} \\
&= \frac{a}{|a|}\frac{c}{|c|}\cdot\rho(x,y).
\end{aligned}$$

110. Puede comprobarse que $\bar{x} = 2$, $\bar{y} = 7$, $\operatorname{cov}(x, y) = 3.6$ y $\operatorname{var}(x) = 2$. Por lo tanto, la recta de regresión lineal es

$$y - 7 = \frac{3.6}{2}(x - 2).$$

O bien, $y = 1.8x + 3.4$. El valor estimado para y cuando $x = 2.5$ es $y = 1.8(2.5) + 3.4 = 7.9$. Se omite la gráfica.

111. Se toman como valores para x los números $1, 2, 4, 5$, y para y los números $4, 8, 18, 22$. Puede comprobarse que $\bar{x} = 3$, $\bar{y} = 13$, $\text{cov}(x, y) = 11.5$ y $\text{var}(x) = 2.5$. Por lo tanto, la recta de regresión lineal es

$$y - 13 = \frac{11.5}{2.5}(x - 3).$$

O bien, $y = 4.6x - 0.8$. El valor estimado para y cuando $x = 3$ es $y = 4.6(3) - 0.8 = 13$. Esta es una estimación del dato faltante en la tabla.

112. El coeficiente principal, esto es, el coeficiente de x, es

$$\frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}} \cdot \sqrt{\frac{\text{var}(y)}{\text{var}(x)}} = \rho(x, y) \cdot \sqrt{\frac{\text{var}(y)}{\text{var}(x)}}.$$

113. A partir de la fórmula general de la ecuación de regresión lineal, resolviendo para x se obtiene la expresión mostrada.

114. Recordemos que la media del conjunto de datos $u = ax + b$ es $a\bar{x} + b$, la media de los datos $v = cy + d$ es $c\bar{y} + d$. Además, $\text{cov}(ax + b, cx + d) = ac \cdot \text{cov}(x, y)$ y $\text{var}(ax + b) = a^2 \cdot \text{var}(x)$. Por lo tanto, la recta de regresión para los datos transformados es

$$v - (c\bar{y} + d) = \frac{ac \cdot \text{cov}(x, y)}{a^2 \cdot \text{var}(x)}(u - (a\bar{x} + b)).$$

O bien,

$$v = \frac{c}{a} \cdot \frac{\text{cov}(x, y)}{\text{var}(x)} (u - (a\bar{x} + b)) + (c\bar{y} + d).$$

115. Los datos x fueron generados en computadora simulando un dado equilibrado. Así es que se espera que tengan aproximadamente las mismas características estadísticas que los datos teóricos y. Esto lo corrobora la gráfica Q-Q elaborada en `R` que se muestra en la Figura A.4.

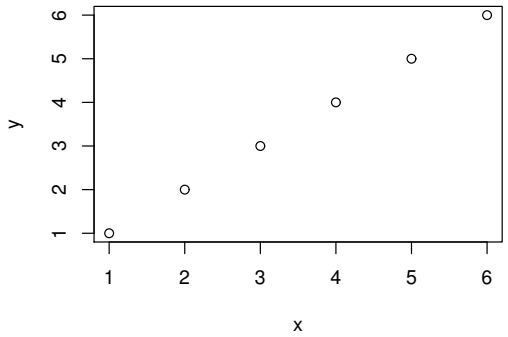


Figura A.4

116. Los dos conjuntos de datos presentan características estadísticas similares: son valores seudoaleatorios de la distribución Poisson de parámetro $\lambda = 5$. La gráfica Q-Q se muestra en la Figura A.5.

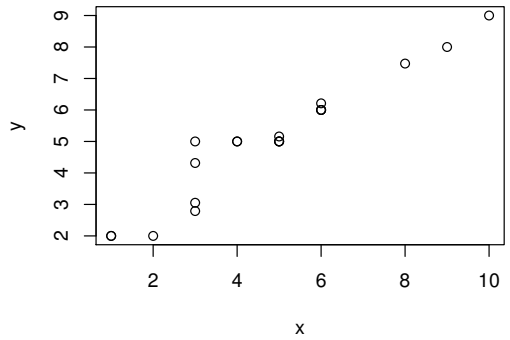


Figura A.5

117. Los datos son valores seudoaleatorios de la distribución normal estándar. La gráfica Q-Q presenta puntos aproximadamente alineados a lo largo de la recta identidad. Véase la Figura A.6.

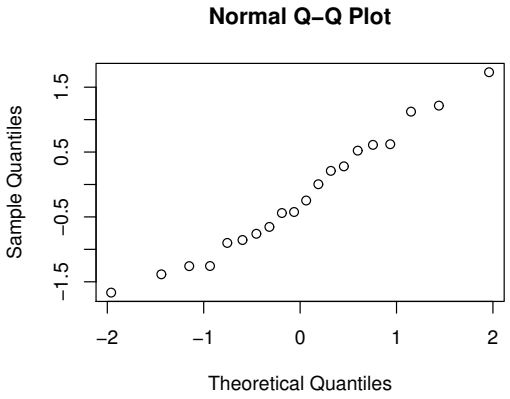


Figura A.6

Bibliografía

[1] Aguirre V., et al. *Fundamentos de probabilidad y estadística.* 2da. edición, Jit Press, 2003.

[2] Alonso Reyes M. del P., Flores Díaz J. A. *Estadística descriptiva para bachillerato.* Serie: Temas de Matemáticas para Bachillerato, Instituto de Matemáticas, UNAM, 2004.

[3] Anderson D. R., Sweeney D. J., Williams T. A. *Statistics for business and economics.* 11th edition, Cengage Learning, 2012.

[4] Berinato S. *Good charts: the HBR guide to making smarter, more persuasive data visualizations.* Harvard Business Review Press, 2016.

[5] Bruce P., Bruce A. *Practical statistics for data scientists.* O'Reilly, 2017.

[6] Chang W. *R graphics cookbook.* O'Reilly, 2013.

[7] Cotton R. *Learning R.* O'Reilly, 2nd. edition, 2013.

[8] Crawley M. J. *The R book.* Wiley, 2012.

[9] Devore J. *Probability and statistics for the engineering and the sciences.* 8th edition, Brooks/Cole Cengage Learning, 2012.

[10] Escobar M. *Análisis gráfico/exploratorio.* La Muralla/Espérides, 1999.

[11] Godfrey M. G., Roebuck E. M., Sherlock A. J. *Concise statistics.* Edward Arnold, 1988.

[12] Hogg R. V., Tanis E. A., Zimmerman D. *Probability and statistical inference.* 9th edition, Pearson, 2013.

[13] Mendenhall W. *Introduction to probability and statistics.* 14th edition, Cengage Learning, 2014.

[14] Sánchez Zuleta C.C., Sepúlveda Murillo F. H. *Estadística descriptiva: exploración de datos con R.* Universidad de Medellín, Colombia, 2015.

[15] Tamayo Vásquez L. G. *Estadística.* Universidad de Medellín, Colombia, 2016.

[16] Tuckey J. W. *Exploratory data analysis.* Pearson, 1977.

[17] Velasco Sotomayor G., Wisniewski P. M. *Probabilidad y estadística para ingeniería y ciencias.* Thomson Learning, 2001.

[18] Vigen, T. *Spurious correlations.* Hachette Books, 2015.

[19] Valencia G., Mendoza M., Aranda F. *Introducción a la inferencia estadística.* Comunicación Interna No. 42, Departamento de Matemáticas, Facultad de Ciencias, UNAM, 1978.

[20] Wexler S., Shaffer J., Cotgreave A. *The big book of dashboards: visualizing your data using real-world business scenarios.* Wiley, 2017.

[21] Walpole R. E., Myers R. H., Myers S. L., Ye K. E. *Probability and statistics for engineers and scientists.* 9th edition, Prentice Hall, 2011.

Índice analítico

R, 161

Agrupamiento de valores, 11
Asimetría
    coeficiente de, 69

Bases de datos, 15
Binomios, 160
Boxplots, 113

Censo, 2
Clase modal, 30
Clases, 12
    marcas de, 12
Coeficiente
    binomial, 160
    de asimetría, 69
        de una v.a., 168
    de correlación, 140
        propiedades, 144
    de variación, 45
        de una v.a., 168
Covarianza, 136
    Propiedades, 140
Cuantiles, 60
    de una v.a., 168
Cuartiles, 62
Curtosis, 72
    curva leptocúrtica, 73
    curva mesocúrtica, 73
    curva platicúrtica, 73
    de una v.a., 168

Datos, 3, 4
    sobre los, 14
Datos agrupados
    descripciones numéricas, 74
Deciles, 62
Descripciones
    gráficas, 97
    numéricas, 23
    para datos conjuntos, 131
Descripciones numéricas
    para datos agrupados, 74
Desviación
    estándar, 40
        de una v.a., 168
    media, 41
        de una v.a., 168
    típica, 40
Diagrama
    de caja y brazos, 113
    de dispersión, 132
    de tallo y hojas, 110
Distribución
    Bernoulli, 165

de probabilidad, 164
normal, 165

Escala de medición, 7
de intervalo, 10
de razón, 10
nominal, 8
ordinal, 8
Esperanza de una v.a., 166
Estadística
descriptiva, 15
inferencial, 16
Exponentes, 158
Extrapolación, 146

Fórmulas
para exponentes, 158
resumen, 77, 151
Factorial, 160
Frecuencias, 49, 51
absolutas, 49
absolutas acumuladas, 53
acumuladas, 53
para datos conjuntos, 134
relativas, 56
relativas acumuladas, 58
relativas porcentuales, 57
Función
de densidad, 165
de distribución, 166
empírica, 117
de probabilidad, 165

Gráfica
de barras, 98
de pastel, 106
de tallo y hojas, 110
Q-Q, 146

Histograma, 102

Intervalo modal, 30
Intervalos, 159

Marca de clase, 12
Media, 24
aritmética, 24
armónica, 83
de una v.a., 166, 168
geométrica, 82
para datos agrupados, 26
Mediana, 30
de una v.a., 168
Medidas
de dispersión, 37
de localización, 23, 24
de tendencia central, 23
Moda, 27
de una v.a., 168
Momentos, 47
centrales, 47
de una v.a., 168
Muestra, 1
tamaño de la, 3

Notación suma, 157

Ojiva, 105
Outliers, 115

Paquete `R`, 161
Percentiles, 63
Población, 1
Polígono
de frecuencias, 103

de frecuencias acumuladas, 104
Porcentajes, 57

Rango, 43
de una v.a., 168
intercuartil, 115
Recta de regresión, 144, 145
Regresión, 144
Resumen de fórmulas, 77, 151
RIC, 115

Skewness, 69
Sumas, 157

Tabla
de contingencias, 134
de frecuencias, 134

Unidad de observación, 2

Valor absoluto, 158
Valores
agrupamiento de, 11
atípicos, 115
Variables, 4
aleatorias, 163
categóricas, 5
clasificación de, 5
continuas, 7
cualitativas, 5
cuantitativas, 5
dicotómicas, 7
discretas, 6
mixtas, 7
Variables aleatorias, 163
continuas, 164
discretas, 164
Varianza, 37
de una v.a., 168
para datos agrupados, 40

Whiskers, 113

Estadística descriptiva
editado por la Facultad de Ciencias de la
Universidad Nacional Autónoma de México,
se terminó de imprimir el 28 de ocubre de 2017
en los talleres de Navegantes de la
Comunicación Gráfica, S.A. de C.V.
Pascual Ortiz Rubio núm. 40, San Simón Ticumac,
Delegación Benito Juárez, C.P. 03660. México, D.F.

El tiraje fue de 1000 ejemplares.

Está impreso en papel book creamy de 60 g.
En su composición se utilizó tipografía
Computer Modern de 11:13.5, 14:16 y 16:18 puntos de pica.

Tipo de impresión: offset.
El cuidado de la edición estuvo a cargo de
Patricia Magaña Rueda.

En este volumen se presentan algunos conceptos elementales de la estadística descriptiva. Esta dirigido a estudiantes universitarios de todas las disciplinas y también a aquellos profesionistas quienes, desde sus muy diversas áreas de conocimiento, requieren conocer alguna técnica o procedimiento para la presentación descriptiva, numérica o gráfica, de la información contenida en un conjunto de datos.

Se hace particular énfasis en mostrar la forma en la que el análisis descriptivo de datos lleva a considerar de una manera más natural los conceptos teóricos que se estudian en cursos de probabilidad y estadística. Se incluyen numerosos elementos gráficos, así como una colección de ejercicios para que puedan ser desarrollados en un salón de clase o bien para proponerlos como tareas a desarrollar en un curso.

Se provee además la respuesta de la mayoría de estos ejercicios para que el lector activo pueda corroborar sus avances.

ISBN: 978-607-02-9724-3