ESTADÍSTICA Y PROBABILIDADES

TOMO 1
NOCIONES BÁSICAS DE ESTADÍSTICA: CONOCIENDO LOS DATOS

JORGE GALBIATI RIESCO

Ministerio de Educación

Gobierno de Chile

UCE
UNIDAD DE CURRÍCULUM Y EVALUACIÓN

ESTADÍSTICA Y PROBABILIDADES.
Tomo 1: Nociones básicas de Estadística. Comprendiendo los datos.
Jorge Mauricio Galbiati Riesco, Ph.D.

MINISTERIO DE EDUCACION
Unidad de Curriculum y Evaluación
Santiago, noviembre 2021

ILUSTRACIONES:
Paola Galbiati Valverde, Diseñadora gráfica.
www.talleronce.cl

EDICIÓN 2021
Este material será distribuido gratuitamente a escuelas públicas del país.

<ins>Portada</ins>:
Calendario maya y números maya.
Los mayas inventaron el cero en forma independiente a su invención en la India, de donde pasó a Arabia y de ahí a Europa. El cero lo representaban por un caracol.

<ins>Agradecimientos</ins>:
A la Doctora en Ciencias Pamela Reyes Santander, Coordinadora de Matemática, Unidad de Currículum y Evaluación, Ministerio de Educación, Gobierno de Chile.
Por la revisión del original y por sus valiosas sugerencias.

Jorge Mauricio Galbiati Riesco, Ph. D.

Doctor en Estadística, Universidad de Iowa, U.S.A

Master en Estadística Matemática, Centro Interamericano de Enseñanza de Estadística (CIENES), Universidad de Chile.
Profesor de Matemáticas y Física y Licenciado en Filosofía y Educación, Universidad Católica de Valparaíso.
Académico de la Pontificia Universidad Católica de Valparaíso desde 1971. Ejerce en el Instituto de Estadística desde su fundación, en 1975. Fue su Director entre el 2001 y 2006.
Ha ejercido la docencia desde 1971, elaborando gran cantidad de material didáctico.
Ha participado en varios proyectos de investigación, con un número de publicaciones en revistas científicas internacionales, fundamentalmente en el tema de Procesamiento Estadístico de Imágenes Digitales.
Ha participado en numerosos proyectos de asesoría estadística a instituciones y empresas.
Entre 2008 hasta 2011 fue miembro del Consejo Superior de la Pontificia Universidad Católica de Valparaíso.
Fue miembro del Consejo Nacional de Estadísticas, entre 2010 y 2011, en representación del Consejo de Rectores de las Universidades Chilenas.
www.jorgegalbiati.cl

ESTADÍSTICA Y PROBABILIDADES

TOMO 1

NOCIONES BÁSICAS DE ESTADÍSTICA: COMPRENDIENDO LOS DATOS

<u>Contenido</u>

Presentación

El currículum escolar chileno en Matemática busca desarrollar las habilidades, actitudes y conocimientos en todos los ejes de la asignatura: números, álgebra y funciones, geometría y probabilidad y estadística. Para comprender las matemáticas y ser capaz de aplicar sus conceptos y procedimientos es necesario trabajar en conjunto con las habilidades de la asignatura: resolver problemas, representar, modelar, argumentar y comunicar, que se desarrollan a lo largo de los doce años de escolaridad, desde el primero básico hasta el cuarto medio.

El espíritu de la Ley General de Educación nos moviliza a proponer y promover de una manera didáctica el aprendizaje de las habilidades y conocimientos matemáticos. También, nos mueve a generar y apoyar iniciativas que tienen como principio el aprendizaje de conocimientos expertos, con contenidos cercanos a los estudiantes y tratados en un lenguaje adecuado para ellos, donde prevalece un objetivo común, que es poner en práctica en las aulas el desarrollo de habilidades y la comprensión del conocimiento para participar con información y de manera propositiva en la vida del país. La Matemática entrega herramientas únicas y poderosas para entender el mundo.

Bajo este gran objetivo, la Unidad de Currículum y Evaluación del Ministerio de Educación de Chile, ha generado recursos pedagógicos y herramientas que promueven el desarrollo de conocimientos y habilidades para el eje de ESTADÍSTICA, entendiendo que los objetivos de aprendizaje de las Bases Curriculares se desarrollan de manera progresiva y en armonía con el nivel etario de los estudiantes. Para los expertos del área de matemática de esta Unidad ministerial, los conocimientos sobre estadística permiten al estudiante recopilar, procesar y organizar datos, también comprender en qué situaciones de la vida diaria es relevante aplicar los conocimientos estadísticos para poder resolver problemas, comunicarse con otros, entender la información de los medios de comunicación e interpretar en base a los datos presentados u obtenidos directamente. En pocas palabras, las Bases Curriculares promueven el pensamiento estadístico.

En este contexto y especialmente dirigido a los niveles de sexto básico a segundo medio, el autor de este libro, en un acto de generosidad para los escolares y profesores de Chile, ha dedicado su tiempo a escribir de manera amable con el lector, y con una sabiduría que se lee desde las primeras páginas, para dejar su legado de muchos años de enseñanza de la estadística, en beneficio no solo de los profesores y estudiantes de nuestro país, sino también de los hispanoparlantes del mundo.

Este libro es una contribución en formato electrónico, que estará disponible en las plataformas digitales del Ministerio de Educación, y representa un gran aporte para la comprensión de las nociones básicas de estadística y su trabajo en aula. El autor ha procurado llevar sus pensamientos de lo que es la estadística y su forma de comprender los conceptos, de manera agrandable y fácil de seguir, ejemplificando y haciendo ver que todo es muy sencillo si se mira de una forma particular. En sus diferentes secciones, muestra el significado de trabajar con las planillas de cálculo para facilitar el trabajo y aprovechar las herramientas disponibles y así poder concentrarse en dar significado a los conceptos y dejar lo más relevante de la información estadística.

Como directora de la Unidad de Currículum y Evaluación del Ministerio de Educación de Chile y además profesora de matemática y física, es un enorme gusto presentarles el título del primer tomo de la serie de **Estadística y Probabilidades "Nociones Básicas de Estadística, comprendiendo los datos"** porque estoy convencida que es y será un aporte significativo para todos los escolares y para las actuales y futuras generaciones de profesores de nuestro país. Los invito a todos a leer y sacar el máximo provecho a todas las ideas aquí presentes, llevar las actividades a sus clases y ajustarlas según sus necesidades contextuales.

Finalmente, aprovecho la oportunidad de ofrecer una sincera felicitación y reconocimiento por el trabajo de investigación y recopilación realizado por el autor del libro, don Jorge Galbiati, y transmitir también el agradecimiento del Ministerio de Educación de Chile por conceder los permisos de publicación de esta obra, de manera gratuita, en la plataforma Currículum Nacional de la Unidad de Currículum y Evaluación.

María Isabel Baeza Errázuriz
Coordinadora Nacional
Unidad de Currículum y Evaluación
MINISTERIO DE EDUCACIÓN

Santiago, noviembre 2021.

Introducción

No hay una definición del término **Estadística** que satisfaga a todos.

Pero algunos dicen que es el arte de **resumir información**, lo que parece una definición muy breve, pero como aproximación, parece ser acertada.

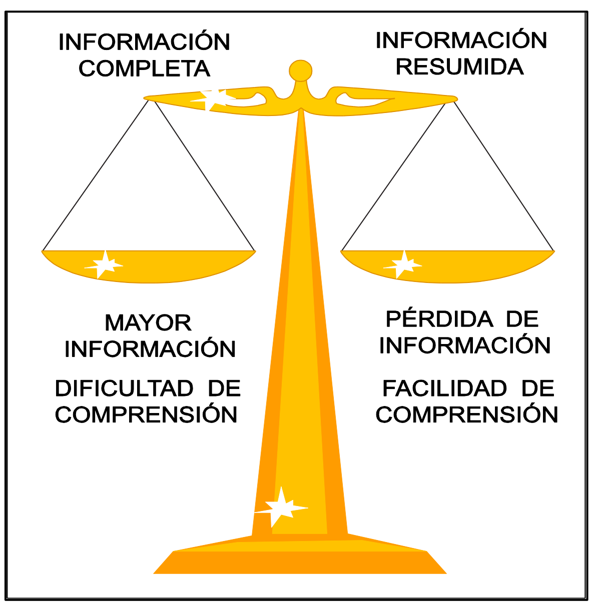
Sin embargo, es muy general, y puede aplicarse también a otras áreas del conocimiento.

¿Para qué resumimos la información que tenemos? Para entenderla mejor, a costa de perder una parte de la **información**.

Los humanos no tenemos capacidad de entender grandes o medianamente grandes volúmenes de información.

Para que la información sea realmente útil, debemos poder entenderla bien, y para eso, es necesario resumirla.

¿Cuánto resumir? Mientras más se resume, más fácil es entenderla, pero hay más pérdida de información.



Hay que buscar un equilibrio entre la facilidad de comprensión que queremos y la cantidad de información que queremos conservar.

EJEMPLO 1

Si tus notas en una asignatura son 5,2; 6,1; 4,8; 6,8; 5,4; 6,2.

El promedio es 34.5/8=5,75

Si alguien te pregunta cómo te está yendo en esa asignatura puedes comunicar tus 6 notas, es decir, responderle que tienes 5,2; 6,1; 4,8; 6,8; 5,4; 6,2.

O bien, simplemente decir que tienes un promedio de 5,75.

En el segundo caso estás dando menos información, pero está comunicando con más claridad el nivel de éxito que tienes en la asignatura.

¿Cómo resume información la Estadística?

La forma más simple es a través de **tablas** y **gráficos**.

Otra forma de resumir es mediante el uso de **medidas** o **indicadores**, que se calculan con los datos. Por ejemplo, el promedio.

Una manera más compleja consiste en desarrollar **modelos estadísticos**, que en términos simples son capaces de describir razonablemente bien una situación real. Son una aproximación de la realidad.

Conociendo el modelo y algunas de sus características, llamadas parámetros, se puede entender mejor el comportamiento del fenómeno real.

La Estadística

La Estadística normalmente extrae información de conjuntos de **datos**, que fueron generados de algún conjunto muy grande de datos, que llamaremos población objetivo o simplemente **población**. Estos conjuntos pueden ser pequeños o muy grandes.

La importancia de los datos en la obtención de información está reflejada en una frase que hemos pedido prestado de quienes se dedican al tema de la **Calidad**. Dice así: "*A Dios le creemos; los demás, que traigan datos*".

De esos datos la Estadística, a través de sus **métodos** propios, es capaz de extraer i**nformación** acerca de la población.

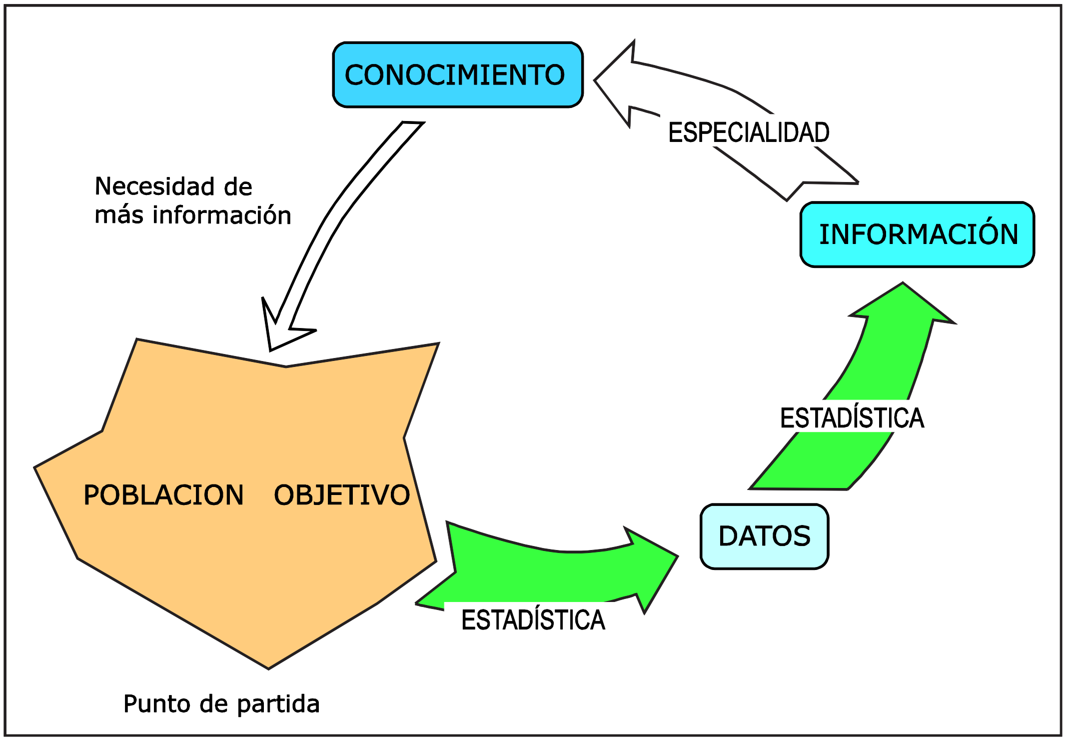
La Estadística también interviene en el proceso de **adquisición** de los datos de donde luego extraerá la información.

Esta información es útil a la hora de tomar **decisiones**.

Quien debe tomar decisiones es un especialista en el tema bajo estudio, lo que él o ella hace es usar sus recursos intelectuales, para agregarle valor a la información y transformarla en **conocimiento** acerca de la población objetivo, que es la información convertida en algo útil.

Este conocimiento sobre la población permitirá reducir la incertidumbre al momento de tomar alguna decisión.

Esto está representado en el siguiente gráfico:



Frecuentemente ese conocimiento conduce a la necesidad de obtener más datos. Que luego aportarán nueva información, que a su vez proveerá más conocimiento.

Y de esa manera el proceso sigue repitiéndose hasta lograr un **conocimiento profundo** del fenómeno sobre el que se deberá **tomar decisiones**.

Las mediciones

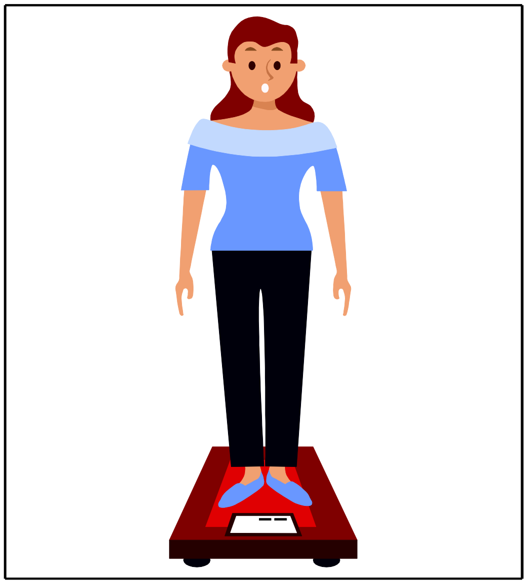
Muchas veces hemos medido algo. Pero ¿qué es medir?

Daremos una definición sencilla de la acción de medir:

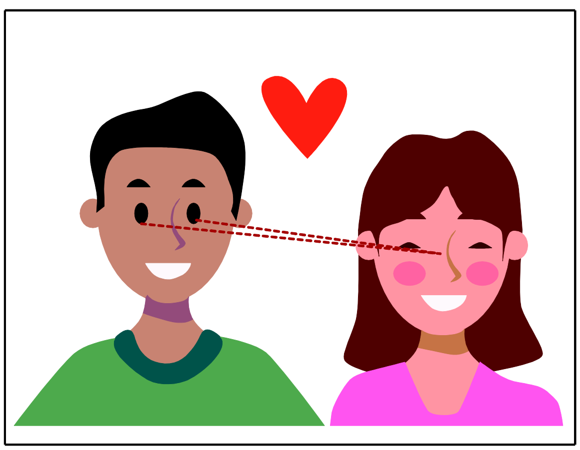
Medir es observar una **característica** y asignarle un **número** o una **categoría**.

EJEMPLO 2

Si subes a una pesa y ves que marca 50 kilos, eso es medir. Asignaste un número al peso.



Si observas el color de ojos de tu compañera y determinas que son verdes, también estás midiendo: asignas una categoría, verde.



En el primer caso usaste una **escala de medida numérica**. En el segundo ejemplo usaste una **escala de medida categórica**.

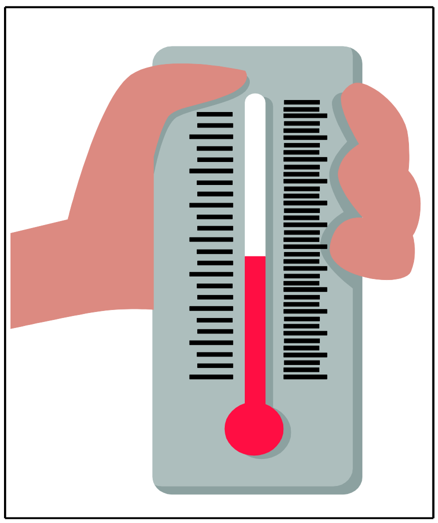
Más adelante nos referiremos a eso.

Las escalas de medida numéricas son mejores que las categóricas, pues nos dan información más precisa sobre la característica que estamos midiendo.

Entre las medidas en escala numérica observamos que hay fenómenos que tienen características con valores positivos y negativos.

EJEMPLO 3

Si se toma la temperatura ambiental seguramente resultarán valores positivos; pero en lugares muy fríos, como Puerto Natales en invierno, pueden resultar temperaturas menores que cero, como, por ejemplo, menos cinco grados centígrados.



Si alguien tiene un emprendimiento y al final del mes calcula sus utilidades, ésta podría ser positiva, como también podría tener la mala suerte de haber obtenido una utilidad negativa, se llama pérdida.

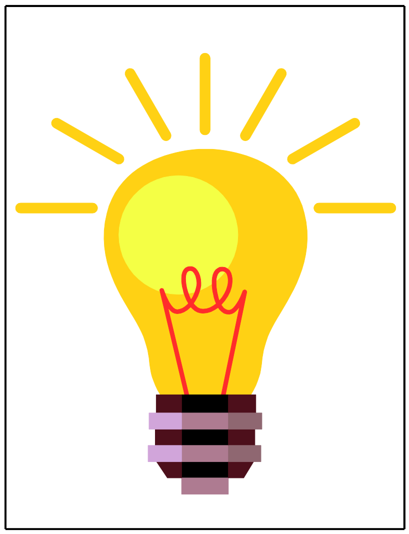
Pero hay otros fenómenos que sólo toman valores positivos. Esto sucede en la mayoría de los casos.

EJEMPLO 4

Algunas características que tienen sólo valores positivos:

Si mides la altura de tus compañeros, su peso, su edad.

El tiempo de vida de una ampolleta, hasta que se quema, la distancia recorrida por un vehículo, el diámetro de unos tornillos, entre muchas otras medidas numéricas.



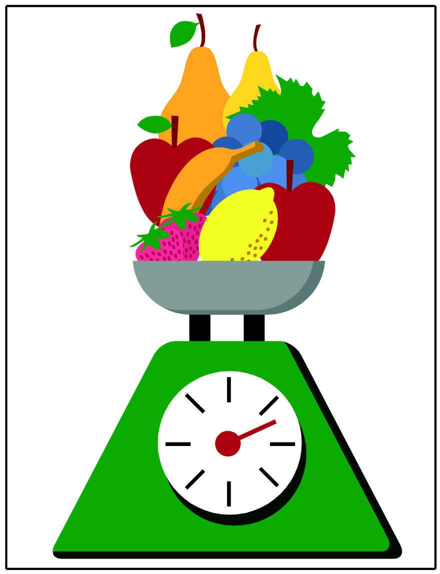
Un **instrumento de medida** es un elemento sensible a lo que se desea medir.

El instrumento permite medir.

EJEMPLO 5

Una regla graduada o una huincha de medir sirve para medir longitudes de objetos.

Un examen médico de laboratorio se usa para medir el contenido de ciertos compuestos químicos en algún órgano de una persona.



Una balanza es para medir el peso de la fruta que estás comprando.

Una encuesta se puede usar para conocer la opinión de un cliente acerca de un servicio que recibió.

La opinión de los jueces en un concurso de saltos ornamentales sirve para medir el desempeño de cada uno de los deportistas que participan en él.

Una prueba de Biología le sirve al profesor para saber cuánto han aprendido sus estudiantes.

Una **escala de medida** es un conjunto de valores con los cuales se expresa una **variable en estudio**.

En una primera clasificación hay dos tipos de escalas de medida: **numéricas** y **categóricas**.

La **numéricas**, como lo dice su nombre, están constituidas por números.

Las **escalas** categóricas están constituidas por objetos que no son números, sino por **categorías**.

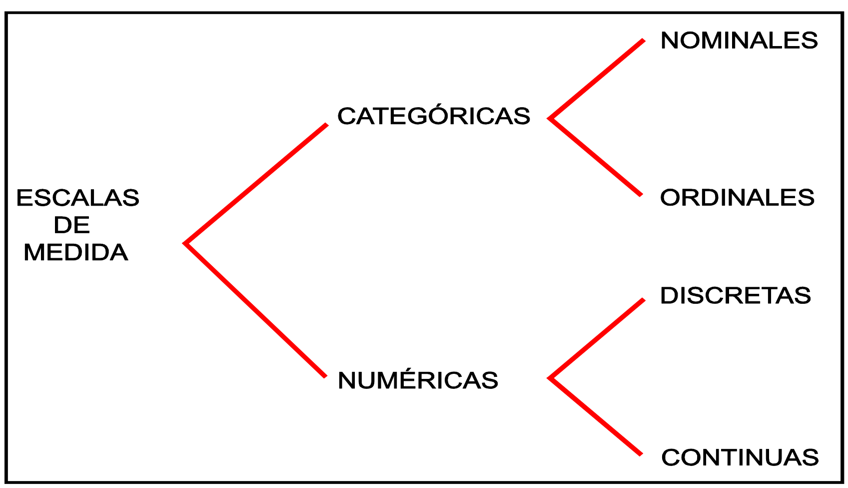
Si una **escala categórica** tiene un **orden natural**, se llama **categórica ordinal**. Si no lo tiene, se llama **categórica nominal**.

Las variables numéricas se pueden dividir en **discretas** y **continuas**.

Las **discretas** son las escalas que son **finitas**, como {1, 2, 3, 4, 5, 6}. Pero también pueden ser **infinitas**, pero sus elementos se pueden **enumerar**: el primer elemento, el segundo, el tercero, …. No tienen límite.

Y las escalas **continuas** son escalas numéricas que son intervalos de números reales. Por eso son **infinitas** y **no se pueden enumerar**.

Este es un concepto bastante **abstracto**. No es fácil imaginarse un conjunto de valores con esa característica.



Si una variable se mide en escala discreta, se dice que es una **variable discreta**.

Si se mide en escala continua, se dice que es una **variable continua**.

Hay casos especiales de escalas, numéricas o categóricas, que sólo tienen dos valores. Se denominan **dicotómicas** o **binarias**.

EJEMPLO 6

Tenemos muchos ejemplos de variables categóricas, como los siguientes conjuntos:

Las marcas de automóviles. En general no tienen orden, **son nominales**.

los colores, que son **nominales**, no hay un orden natural.

los meses del año, que son ordinales porque tienen un **orden**.

{bueno, regular, malo}; esta escala tiene un orden natural, es **ordinal**.

{alto, mediano, bajo}; tiene orden, es **ordinal**.

El orden alfabético es arbitrario, así que estrictamente no dan un orden natural.

Eso es porque el orden en que están las letras del alfabeto se lo dieron arbitrariamente quienes lo crearon.

Puede que en otro alfabeto distinto al que estamos usando aquí, el orden sea diferente.

Ejemplos de conjuntos de medidas **numéricos finitos** pueden ser {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, o bien {3; -6; -3,9; 25,8}. En este último separamos los números con punto y coma, para que no haya confusión con la coma decimal.

Las escalas numéricas continuas tienen la siguiente particularidad: si tomamos un número cualquiera del conjunto, **no hay un número siguiente**.

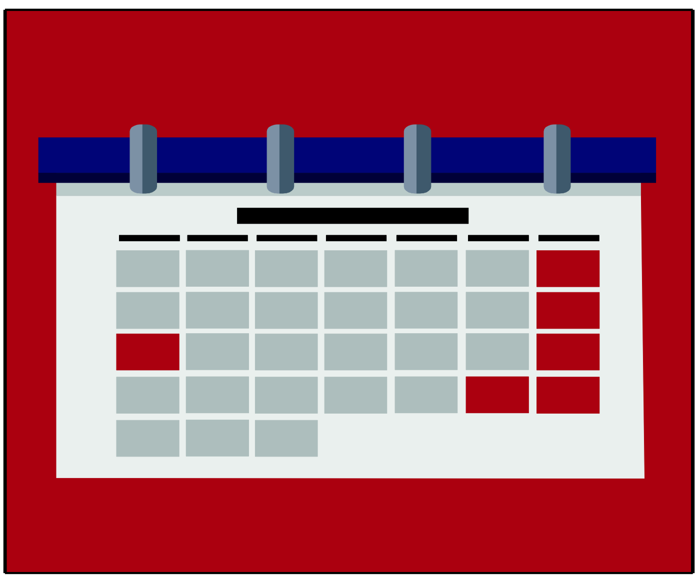
Por ejemplo, tomemos el 3,456. ¿Cuál es el siguiente? No tiene. Si alguien dice que es el 3,457, resulta que no es el siguiente. Porque antes de este último hay muchos, como el 3,4565 entre otros.

Cada vez que tengamos dos números de una escala **continua**, podremos encontrar uno que está entre ellos. En realidad, hay **infinitos** números reales entre el 3,456 y el 3,457. Y no se pueden enumerar.

Considera, por ejemplo, un número real muy especial, **pi**, que seguramente conoces. Es el cociente entre el perímetro de un círculo dividido por su diámetro. Pi= 3,14192653589793….. **¡tiene infinitos decimales!**

EJEMPLO 7

Si observas el mes de nacimiento de un grupo de personas, tienen orden, por lo que estás usando una escala de medida categórica ordinal.



Si registras la actividad laboral de las mismas personas, como arquitecto, jornalero, suplementero, enfermera, vendedora, etc., no tienen un orden natural.

El orden alfabético es arbitrario; en otras culturas el orden de las letras, que representan sonidos emitidos por la boca, es distinto.

EJERCICIOS

1) En un condominio se contó el número de niños y niñas menores de 13 años por vivienda.

Indica y clasifica el tipo de la variable que se está midiendo: si es numérica discreta, numérica continua o categórica.

2) En cada caso indica una variable que mida alguna característica de los estudiantes de tu establecimiento educacional, y que se clasifiquen como:

a) numérica continua

b) numérica discreta

c) categórica.

Poblaciones y muestras

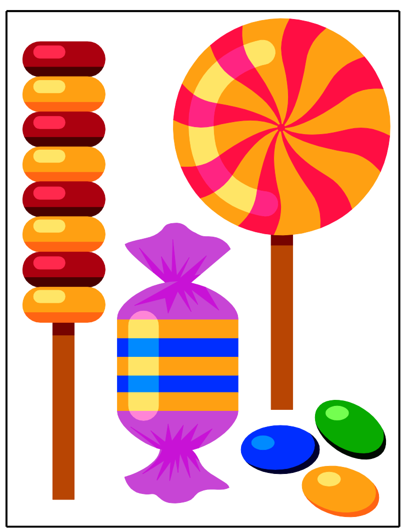
Siempre que queremos estudiar algo y descubrir qué propiedades tiene, nuestras conclusiones se refieren a una parte de la realidad que nos rodea.

EJEMPLO 8

Si quieres averiguar qué piensan tus compañeros acerca de hacer un paseo el sábado próximo, lo que logres averiguar se referirá a la totalidad de tus compañeros. Pero sólo le preguntarás a algunos.

Si una fábrica de caramelos quiere saber cuál será la aceptación que tendrá un nuevo tipo de bombón que quiere introducir, llevará a cabo un estudio de mercado, y los resultados que obtenga se referirán al público a quienes querrá vender su nuevo producto.

Pero no le puede preguntar a todos los posibles consumidores de bombones.

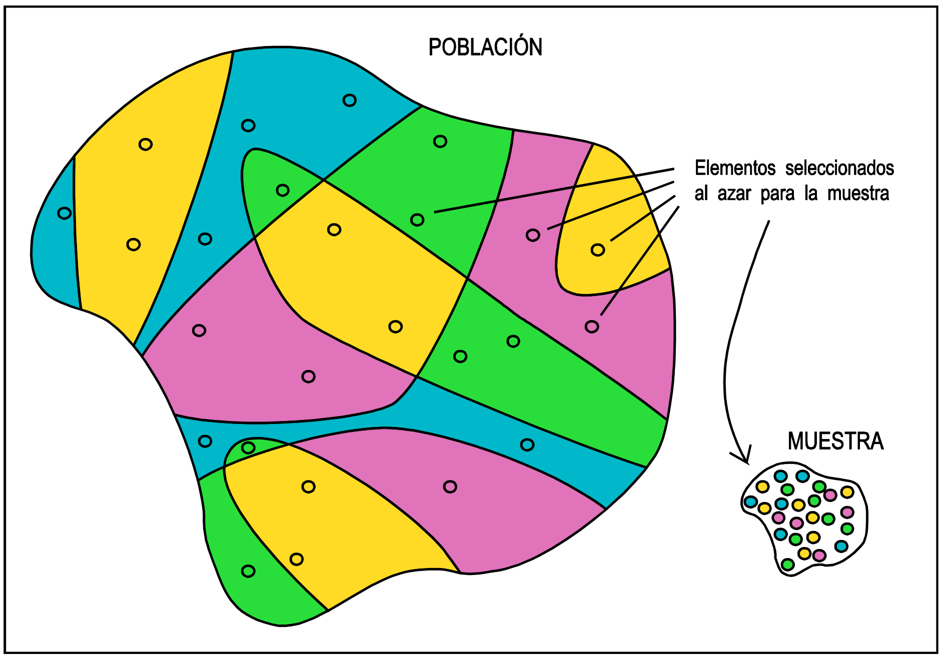


Como dijimos antes, el conjunto de todas las personas o todas las cosas a las cuales se refiere un estudio, se denomina **población** o bien **población objetivo**.

Las poblaciones suelen ser **muy grandes**, por ejemplo, los consumidores, en todo el país, de una determinada marca de jabón.

Por eso, para hacer un estudio, se suele observar sólo una fracción de la población. Esta se llama **muestra**.

La muestra, si es **representativa**, puede considerarse como una **miniatura** de la población.



Para que el estudio de una población, a través de una muestra, sea verdaderamente válido, la muestra debe ser **representativa** de la población.

Una condición que debe cumplir es que sea seleccionada **aleatoriamente**, o **al azar**.

Estrictamente, las poblaciones y las muestras están formadas por las **medidas** de la variable en estudio y no por los **sujetos** que son medidos.

Sin embargo, en este contexto consideraremos las poblaciones y las muestras constituidas por los sujetos.

La característica de la población que medimos, porque es la que nos interesa estudiar, se llama **variable en estudio**.

Por ejemplo, si en un momento dado se mide la temperatura de un compuesto en una reacción química, la variable en estudio es la temperatura que se observa.

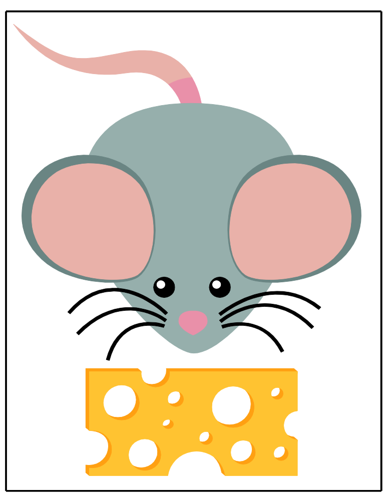
Otra situación puede ser la siguiente: se asignan temas a los estudiantes de un curso para que escriban composiciones sobre ellos; el profesor quiere medir el **grado de creatividad** de los estudiantes.

Sea como fuere la forma como el profesor la medirá, es la **creatividad** lo que le interesa. Esa es la **variable en estudio**.

La **variable en estudio** se puede medir en cualquiera de las escalas de medida que introdujimos antes: numérica, discreta o continua, o bien categórica, nominal u ordinal.

EJEMPLO 9

Un investigador examina los efectos de un agente patológico presente en un determinado alimento, en ratones.



Tiempo después, durante el cual los ratones consumieron dicho alimento, el investigador los examina para detectar presencia o ausencia de posibles indicios de la patología.

Identificaremos la población objetivo, la muestra, la unidad experimental, la variable en estudio, el tipo de variable y la escala de medida.

La población objetivo es el conjunto de los ratones del tipo a que se refiere el resultado del estudio.

La muestra es el grupo de ratones usados en el experimento.

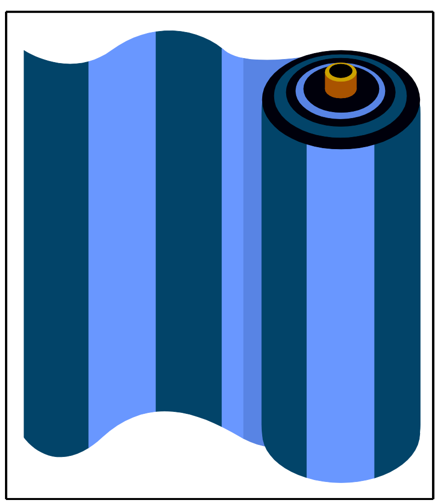
La variable en estudio es la presencia o ausencia de la patología.

Los datos son las medidas tomadas a los ratones. En este caso se mide si tienen o no tienen la particular patología.

Es una variable **categórica**, es decir, **no numérica**, tiene dos valores posibles: tiene o no ella patología. Este tipo de variable con dos valores suele llamarse **dicotómica**.

EJEMPLO 10

Un inspector de control de calidad en una industria textil selecciona y examina, en varios momentos de un día, piezas de género producidos, pesando la pieza y calculando el peso por metro.



Le interesa verificar que sea lo más uniforme posible.

La **población objetivo** consiste en todas las piezas de género similares producidas en esa industria.

La **muestra** consiste en todas las piezas seleccionadas y examinadas por el inspector de calidad.

La **variable en estudio** es lo que mide el inspector para verificar la calidad, el peso por metro.

En este caso la variable en estudio es el peso por metro, medida en escala numérica, y los datos son los valores numéricos de las medidas obtenidas por el inspector de control de calidad.

EJERCICIOS

3) Se desea saber si un sistema nuevo de enseñanza de Lenguaje tiene un efecto beneficioso sobre el aprendizaje de los alumnos.

Se aplica este nuevo sistema durante un año escolar en 10 cursos seleccionados por sorteo, separadamente en escuelas básicas de la comuna San Fabián.

Al final de año se aplicará una evaluación. Identifica:

a) la población objetivo

b) la o las variables involucradas en el estudio.

4) Se desea estimar el gasto promedio, por vivienda, en combustible durante los últimos tres meses, en una ciudad de tamaño pequeño.

Para ello se seleccionan 362 viviendas, al azar, y en cada una se pregunta por el gasto en combustible. Se sabe que en total hay 27.640 viviendas.

Identifica:

a) la variable que se interesa medir

b) la población objetivo.

5) Indica dos variables que midan alguna característica de un estudiante típico de tu colegio, y que se clasifiquen como:

a) continua

b) discreta

c) categórica.

6) Identifica y clasifica las variables involucradas en los siguientes estudios:

   a) Estructura por sexo y edad de la población del país

   b) número de hijos por familia en las provincias de un país

   c) mortalidad por causa de muerte

   d) proporción de individuos de una población diagnosticados como extrovertidos, en un estudio psicológico

   e) tiempo, en centésimas de segundo, entre el inicio de la llegada de dos mensajes consecutivos en un sistema de comunicación electrónico

   f) palabras leídas en 15 segundos, por una muestra de individuos.

Selección al azar con y sin reposición

MOTIVACIÓN

En las encuestas de opinión se supone que la selección de las personas a las que se les pide que respondan la encuesta, a través de un cuestionario, se hace **al azar** y **sin reposición**.

Sin reposición significa que, si alguna persona respondió la encuesta, a ella no se le pedirá otra vez que responda el mismo cuestionario.

Ya no volverá a salir seleccionada en la muestra.

Hay varias formas de elegir una muestra aleatoria; distinguimos dos de ellas: selección **con reposición** y **selección sin reposición**.

Selección de una muestra al azar **con reposición** significa que si sale seleccionado un elemento de la población, podría volver a ser seleccionado, si por el azar eso sucediera.

Mientras que en la elección de una muestra al azar **sin reposición**, una vez que un elemento es seleccionado, ya no puede serlo otra vez.

EJEMPLO 11

Supongamos que tenemos la siguiente población, consistente en las alturas, en centímetros (cm), de un grupo de 30 alumnos de Primer Año Medio:

146	175	147	143	170	172	177	162	160	164
185	151	180	161	152	171	182	163	181	169
148	184	166	157	176	156	149	178	153	168

Esta es nuestra población objetivo.

Si obtenemos el promedio de estos valores, veremos que el promedio es 4946/30=164,8667 cm.

Nunca debemos olvidar de agregar la unidad en que se mide lo que queremos expresar.

Hay excepciones, que se producen cuando las observaciones no tienen unidad, como, por ejemplo, cocientes entre alto y ancho de algún objeto.

En todos los demás casos, **debe estar la unidad de medida**.

Recordemos que el promedio es una medida de **centro** o de **tendencia central**. Es como un **representante** de la población.

Supongamos que obtenemos una muestra de 10 alumnos, seleccionados totalmente al azar, sin reposición, y resultó ser la siguiente:

Muestra 1	147	163	184	143	172	177	151	169	181	164

El promedio de la muestra es $165,1m. Se puede ver que es parecido (pero no igual) al promedio de toda la población, $164,8667cm.

Ahora obtendremos tres muestras aleatorias más. La muestra 2, sin reposición.

Las muestras 3 y 4 son con reposición. En la 3, se repitió el 161. En la 4 se repitió el 171.

muestra2	176	178	175	170	185	147	149	171	169	153
muestra3	146	170	169	182	185	185	166	156	169	163
muestra4	163	178	161	176	178	184	151	161	181	182

Los promedios de las muestras 2, 3 y 4 son, respectivamente, 167,3; 169,1 y 171,5 cm. Todos parecidos al verdadero promedio de la población, 165,8667cm.

Una pregunta sobre la cual podemos reflexionar es: ¿son suficientemente cercanos los promedios de las muestras al promedio de la población objetivo?

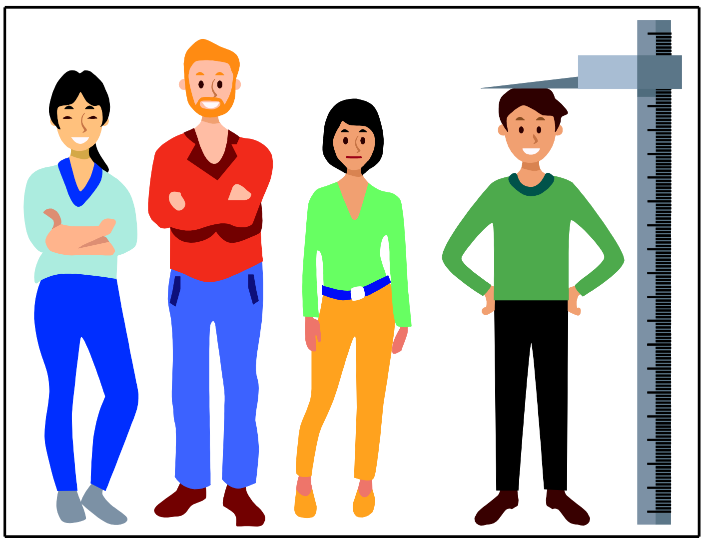
ACTIVIDAD PRÁCTICA

Haz una lista de las alturas de todos los alumnos del curso.

Calcula el promedio.

Corta un papel en cuadritos, escribe la altura en cada uno, dóblalos y ponlos en una caja o bolsa.

Fija un tamaño de muestra, por ejemplo, 7, y obtén una muestra al azar, sacando papelitos de la bolsa. Calcula el promedio de las alturas de la muestra y compáralo con el promedio de la población.



Repite lo anterior varias veces, con y sin reposición.

ACTIVIDAD COMPUTACIONAL EXCEL

Usando el sistema Excel, ingresa los valores de las alturas de todos tus compañeros incluyendo la tuya. Esta es tu población.

Calcula el promedio de las alturas usando la función.

Obtén muestras aleatorias de distintos tamaños, con y sin reposición. Para esto deberás ir a **Datos** y abrir el menú **Análisis de datos**.

Elige **muestra**. Aparecerá un menú en que deberás ingresar el rango de los datos que ingresaste, el número de muestras (el tamaño de la muestra) y el rango donde quieres que esté la muestra. Repite lo anterior varias veces.

En cada caso calcula el promedio y compárelo con la altura promedio poblacional.

La muestra es con reposición.

La variación

MOTIVACIÓN

Si compramos cerámica y pasado un tiempo nos falta y compramos más, de la misma marca y tipo, es muy posible que nos toque de otra partida de producción, y el color sea ligeramente distinto.

Todos los fenómenos que observamos tienen una propiedad común: tienen variación; es decir, si el fenómeno se repite, sus características serán ligeramente diferentes.

Esta variación se debe a **múltiples factores** que influyen en **pequeña medida**, como cambios en la temperatura, corrientes de aire, estados de ánimo de las personas involucradas, entre muchas otras.

Algunos fenómenos tienen variación muy pequeña, como los **experimentos de laboratorio**.

Otros, como los **procesos industriales**, tienen una variación moderada.

Y hay otros, en los que está muy **involucrado el ser humano**, que tienen mucha variación. Estos son los fenómenos sicológicos, económicos, sociológicos, educacionales, políticos, entre otros.

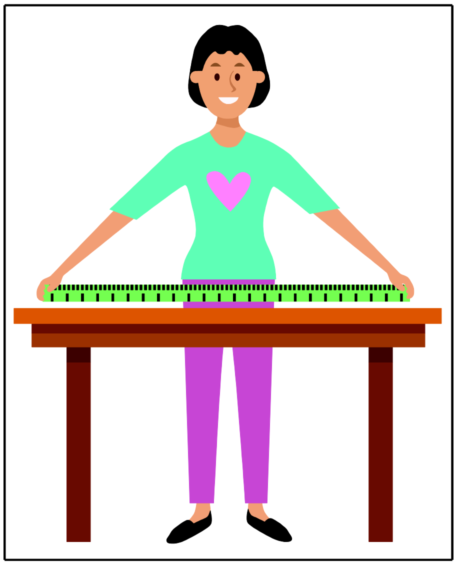
Si se efectúan varias mediciones sobre un mismo objeto, éstas pueden presentar variación, sobre todo si interviene el ser humano en el proceso de medición.

ACTIVIDAD PRÁCTICA

Toma una regla graduada.

Que cada alumno mida el ancho de la mesa del profesor, con precisión de hasta medio milímetro, y lo anote en un papel, sin que los demás sepan el valor que obtuvo.

Al final, hacer la lista de todas las medidas obtenidas.



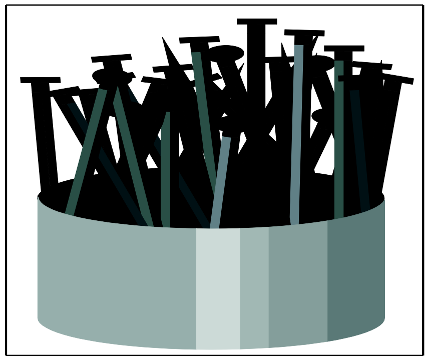
Se puede observar que hay variación entre los valores observados. Aquí la variación es entre las mediciones.

Seguramente muchos valores son parecidos o iguales, mientras que hay unos pocos valores extremos, que se apartan de la mayoría.

ACTIVIDAD PRÁCTICA

Toma una cantidad de clavos de una misma medida, por lo menos de 2 pulgadas, digamos que al menos 30 clavos.

Con una regla graduada en milímetros, mide y registra la longitud de cada clavo.



Verás que no todos miden lo mismo, incluso cuando proceden de un mismo paquete.

Con los datos obtenidos, construye un gráfico o un histograma. ¿Es apropiado, o sería mejor un gráfico de barras?

Describe lo que observas.

EJERCICIOS

Se tiene una población muy pequeña, formada por los números 2, 5, 7, 9, y 12.

a) Extrae todas las muestras posibles de 2 elementos, sin reposición. ¿Cuántas son?
   Por cada muestra calcula el promedio y compáralo con el promedio de toda la población. Verás que algunas dan parecido y otras dan muy distinto.
b) Extrae todas las muestras posibles de 2 elementos, con reposición. ¿Cuántas son?
   Por cada muestra calcula el promedio y compáralo con el promedio de toda la población. Ahora hay mayor variedad de valores para los promedios de las muestras, algunos muy distintos del promedio de la población.

En una industria, una máquina automática de embotellado de jugo en cajas tiene detenciones por cajas que ingresan en mala posición y traban el sistema.

Los siguientes datos corresponden a la cantidad de detenciones de la máquina automática ocurridos durante los primeros 60 días de uso, hasta que se decidió cambiarla.

Día	Detenciones	Día	Detenciones	Día	Detenciones
1	8	21	2	4	5
2	1	22	5	42	1
3	4	23	8	43	8
4	0	24	0	44	5
5	5	25	7	45	2
6	2	26	0	46	5
7	4	27	4	47	4
8	2	28	2	48	3
9	6	29	5	49	4
10	2	30	5	50	2
11	3	31	2	51	2
12	1	32	4	52	0
13	1	33	5	53	5
14	3	34	0	54	1
15	5	35	4	55	3
16	6	36	3	56	1
17	1	37	0	57	3
18	3	38	5	58	4
19	0	39	6	59	1
20	8	40	3	60	5

Asumamos que esta es nuestra población. En promedio son 3 detenciones diarias.

Obtén una muestra aleatoria.

Antes deberás decidir el tamaño de la muestra, y si va a ser con o sin reposición.

También deberás decidir el método para asegurarte que la muestra sea tomada al azar. Por ejemplo, enumera papelitos con los números del 1 al 60 y extrae papeles en forma aleatoria.

Calcula el promedio del número de detenciones en la muestra y compáralo con el valor de la población.

Una advertencia: está comprobado que las personas no funcionan bien como generadores de procesos aleatorios.

Si tú tomas la decisión sobre cuáles estarán en la muestra, éste resulta no ser un buen método de selección de una muestra al azar.

Gráficos de barra e histogramas

MOTIVACIÓN

Una imagen dice más que mil palabras. Este es un dicho que seguramente conoces. Pero es muy cierto.

Por eso la Estadística ha entendido que los gráficos son importantes para resumir y transmitir información.

Siempre que estén bien hechos, **no distorsionen** la información, y sean **autocontenidos**.

Esto último significa que debieran explicarse por sí solos, sin necesidad de tener que buscar en el texto cómo interpretar o qué se pretende mostrar con el gráfico.

En esta sección explicaremos lo relacionado con dos gráficos muy utilizados, que se parecen, pero no son iguales: el **gráfico de barras** y el **histograma**.

Primero consideraremos datos de tipo categóricos, o sea, **no numéricos**.

Hay básicamente dos formas de **presentar** estos datos.

Una es como una lista en que cada uno aparece como una observación individual, con los valores de la categoría a que pertenece.

Estos pueden estar en algún orden determinado, o simplemente desordenados.

A esta manera de presentar los datos se le suele referir como **datos a granel**.

Una segunda forma de presentar un conjunto de datos categóricos, de tal modo que sea más entendible, es mediante una **tabla de frecuencias**.

Esta tabla tiene, en su forma más simple, dos columnas, en una están las **categorías** y en la otra están las **frecuencias**, que son las veces que aparece cada categoría.

Si los datos están **a granel**, lo recomendable es **tabularlos**, o sea, construir una **tabla de frecuencias** a partir de ellos.

Para ello se clasifica cada dato en la categoría que le corresponde, y se lleva la cuenta. La frecuencia de una categoría es el número de datos que tienen esa categoría.

Con eso no estamos resumiendo la información, sino que estamos presentando los datos de una forma más fácil de entender.

El paso siguiente es representar las frecuencias en un **gráfico de barras**, que, como lo dice su nombre, consiste en barras, verticales u horizontales, cuyas longitudes son proporcionales a las frecuencias.

Veamos estas cosas en el siguiente Ejemplo con datos categóricos.

EJEMPLO 12

Supongamos que en el año 2019 egresaron 982 personas del Instituto Profesional Aurora, en las especialidades de Administración, Literatura, Educación, Biología y Ciencias Sociales.

De los 982 egresados, 512 son hombres y 470 son mujeres.

La lista completa, en orden alfabético de apellidos, podría tener una forma como la siguiente (seguramente incluiría otros datos adicionales)

Número de orden	Nombre	Apellido	Especialidad
1	Patricio	Abarzúa	Literatura
2	Constanza	Alvarez	Biología
3	Catalina	Ascui	Educación
…	…	…	…
981	Benjamín	Zamora	Administración
982	Javier	Zárate	Ciencias Sociales

A partir de esta lista, contando cuántos casos hay por cada especialidad, y por sexo, podemos elaborar una **tabla de doble entrada**, que muestra el número de Hombres, de Mujeres y los totales egresados, en cada Especialidad.

Es la siguiente:

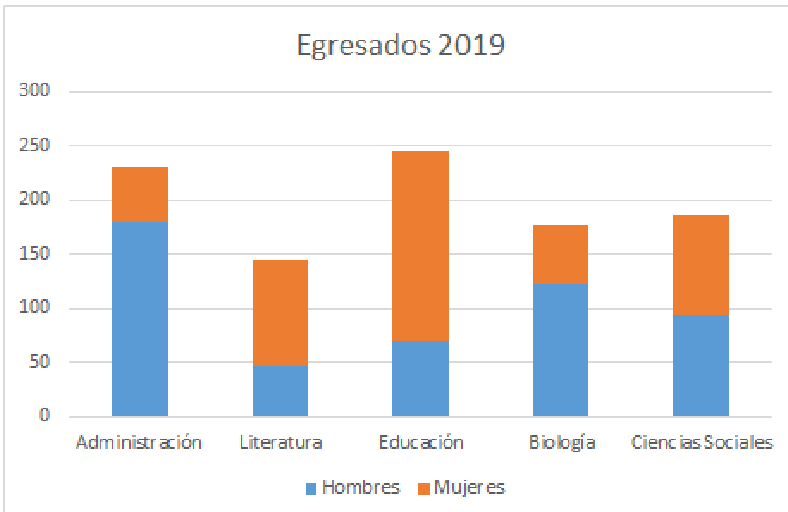
Especialidad	Hombres	Mujeres	Total
Administración	180	50	230
Literatura	46	98	144
Educación	70	175	245
Biología	122	55	177
Ciencias Sociales	94	92	186
Total	512	470	982

En esta tabla se ha **resumido información**, pues no aparecen individualizadas las personas, pero entrega un panorama mucho **más claro** de cómo se distribuyen los egresados según especialidad y sexo.

El siguiente paso, para presentar estos datos con mayor claridad, es construir un **gráfico de barras**. Para ello usaremos Excel.

Para distinguir entre el número de Hombres y de Mujeres, las barras las dividiremos en dos trozos de colores diferentes, siendo la longitud del trozo de abajo proporcional a la frecuencia de los Hombres, la longitud del trozo de arriba proporcional a la frecuencia de las Mujeres, por cada una de las especialidades.

El resultado es el siguiente:



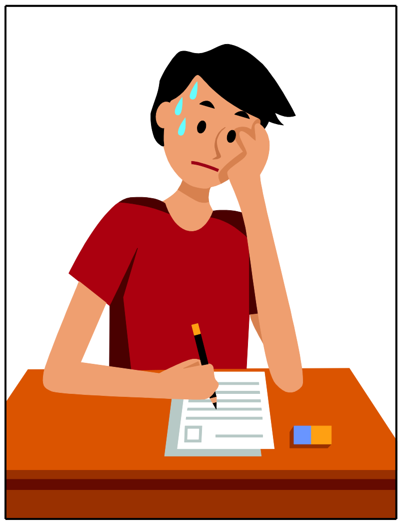
Observa que el orden en que se ponen las barras es arbitrario. Se podría haber puesto Ciencias Sociales primero, sin que esto nos ocasione problemas.

Esto es porque representan categorías, sin orden.

Ahora vamos a ver otro ejemplo, con datos numéricos en lugar de categorías.

EJEMPLO 13

Los datos que se presentan a continuación son los puntajes obtenidos en un examen por un grupo de 25 postulantes a un trabajo (en escala de 1 a 20 puntos).

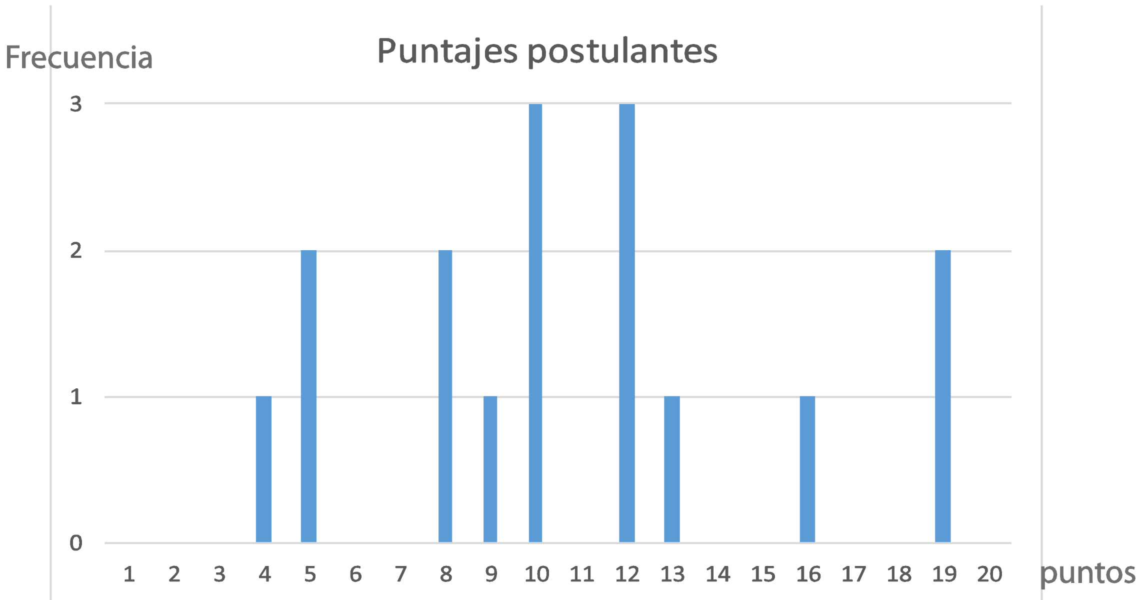


10	4	5	10	15
12	19	13	16	6
8	5	12	11	15
19	8	12	11	7
16	10	9	13	13

Son datos a granel. Los vamos a convertir en una tabla de frecuencias:

Puntajes	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Frecuencias	1	2	1	1	2	1	3	2	3	3	0	2	2	0	0	2

Vamos a representar estos datos en un gráfico de barras, nuevamente utilizando Excel.



Observa que las barras están ordenadas de acuerdo con el valor numérico que representan.

En el Ejemplo anterior, no representan números, y no tienen definido un orden.

En este caso son números. Sería poco afortunado alterar el orden natural de los números.

Ahora explicaremos qué es un **histograma**, sus similitudes y sus diferencias con un gráfico de barras.

En primer lugar, se debe construir una **tabla de frecuencias por intervalos**.

Se definen intervalos, y se cuenta el número de los datos que caen en cada uno de esos intervalos.

A continuación, describiremos el procedimiento para construir los intervalos. Parece complicado, pero en realidad es bien simple.

Para construir los **intervalos**, primero se debe decidir **cuántos** intervalos queremos.

Si hay **muchos** intervalos, se resumirá poco la información y el resultado será **poco entendible**.

Si hay muy **pocos** intervalos, se resume demasiado, perdiéndose mucha de la información.

Coma ya hemos dicho antes, hay que buscar un punto intermedio entre **resumir** y **no perder demasiada información**.

Una receta simple, si tenemos *n* datos, es extraer la raíz cuadrada de *n* y aproximarla al entero.

Ese puede ser un punto de partida para tomar la decisión sobre el número de datos, aumentándolo o disminuyéndolo según lo que uno estime conveniente.

De esa forma, como una pauta, si hay 30 observaciones, pensar entre 5 o 6 intervalos; con 50 observaciones, 7 intervalos; 80 observaciones, 8 o 9 intervalos; 150 observaciones, 12 o 13 intervalos; 200 observaciones, 14 o 15 intervalos, etc. Pero esto es sólo una pauta.

Para definir los intervalos, considerar la menor y la mayor de las observaciones.

Si la **diferencia** entre ambos valores es D y el **número de intervalos** es k, el ancho de cada intervalo es

$$\frac{D}{K}$$

Resultado que se aproxima para que sea un número *entero*.

Se elige un punto de partida de modo que no sea mayor que el menor de los datos, para que éste no quede fuera de los intervalos.

Hay que tener cuidado de que el mayor de los datos no sea mayor que el límite superior del último de los intervalos.

Se esto pasa, se debe aumentar el ancho de los intervalos o bien aumentar el número intervalos. No pueden quedar datos fuera de los intervalos.

Una vez definidos los intervalos, se procede a **clasificar** cada dato en el intervalo que lo contiene, y llevar la cuenta.

Antes, hay que decidir qué hacer si una observación **coincide con el límite** entre dos intervalos: si se incluirá en el intervalo inferior o en el intervalo superior.

Esa decisión es arbitraria, pero una vez tomada, debe aplicarse a todos los datos por igual.

Terminada la **tabla de frecuencias por intervalos**, procedemos a construir el histograma:

El histograma consta de un eje horizontal que representa los números reales.

En este eje se trazan marcas que corresponden a los límites de los intervalos previamente definidos.

Se dibujan rectángulos, cuyas bases son los intervalos, y cuyas alturas son proporcionales a las frecuencias.

De este modo, los rectángulos correspondientes a intervalos seguidos quedan pegados uno al lado del otro.

En el gráfico de barras, éstas están separadas.

EJEMPLO 14

El dueño de un mini supermercado quiere conocer cómo son sus ventas semanales de bebidas.

Para eso registró las ventas durante 30 semanas seleccionadas al azar, de entre los dos últimos años.

Los datos que obtuvo son los siguientes, en miles de pesos:

93,4	68,3	32,4	104,8	104,4	92,0
82,8	76,5	112,2	85,7	115,2	106,4
80,6	79,8	89,6	98,0	37,9	52,1
93,2	62,2	66,1	65,2	42,1	88,4
110,0	78,5	59,7	117,1	109,5	59,2

Haremos una **tabla de frecuencias por intervalos** con estos datos.

Son 30 datos, por lo tanto 5 puede ser un buen número de intervalos.

El valor menor es 32,4 y el mayor es 117,1 mil pesos, la diferencia es 84,7 mil pesos.

84,7/5=16,94, podemos redondearlo convenientemente a 20.

Entonces definiremos cinco intervalos de 20 mil pesos cada uno, partiendo de 20 mil y llegando hasta 120 mil.

Con esto nos aseguramos de que todos los valores que aparecen en la tabla estén incluidos en los intervalos.

Antes de contar debemos tomar una decisión: Si algún valor llegara a coincidir con el límite de dos intervalos, ¿en cuál de los dos lo incluiremos?

Esta es una decisión arbitraria que se aplicará en todos los intervalos; en este Ejemplo decidiremos ponerlo en el intervalo superior.

Los intervalos y las frecuencias en cada uno se muestran a continuación.

Se ha incluido una columna para ir contando los valores que están en cada uno de los intervalos, algo que hay que hacer con mucho cuidado para no dejar ningún valor afuera.

Intervalo	Conteo	Frecuencia
20 a 40	//	2
40 a 60	////	4
60 a 80	///////	7
80 a 100	/////////	9
100 a 120	////////	8

Con esta tabla construiremos un histograma.

En el gráfico de barras cada valor es un punto en el eje horizontal.

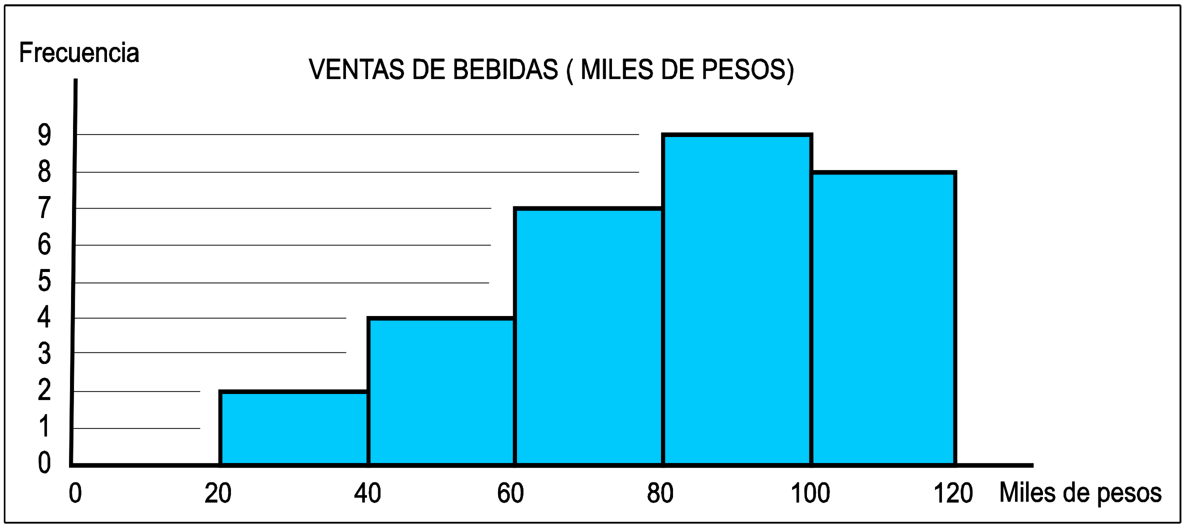
Sólo los valores que aparecen en la muestra están representados, y no hay más.

El histograma es diferente. El eje de horizontal representa un **continuo** de números, los **números reales**. Cada punto del eje representa un número real, un posible valor de la muestra.

Las frecuencias están representadas por rectángulos, cuyos lados coinciden con los límites de los intervalos.

Por lo tanto, los rectángulos correspondientes a intervalos contiguos se representan **pegados** uno al lado del otro.

El histograma se construyó con un programa para dibujar.



Una **observación**: En un gráfico de barras, las frecuencias son proporcionales a las **longitudes** de las barras.

En un histograma, las frecuencias son proporcionales a las **áreas** de los rectángulos.

Pero como en la gran mayoría de los casos, y **siempre para nosotros**, los intervalos son de **igual longitud**, las bases de los rectángulos son iguales.

En consecuencia, cuando los intervalos son iguales, las alturas de los rectángulos **resultan** proporcionales a las **frecuencias**, como en los gráficos de barras.

Debes tener cuidado, porque en la literatura a veces usan la palabra histograma para referirse a un **gráfico de barras**.

Una **precaución** que debes tener es la siguiente:

A veces en la literatura se presentan gráficos muy hermosos y llamativos, que suelen llamar **infogramas**.

Hay que tener cuidado, porque estos tienden a distorsionar la información, o a ser poco claros en lo que quieren transmitir.

Esto puede ser aceptable en algunos contextos **informales**.

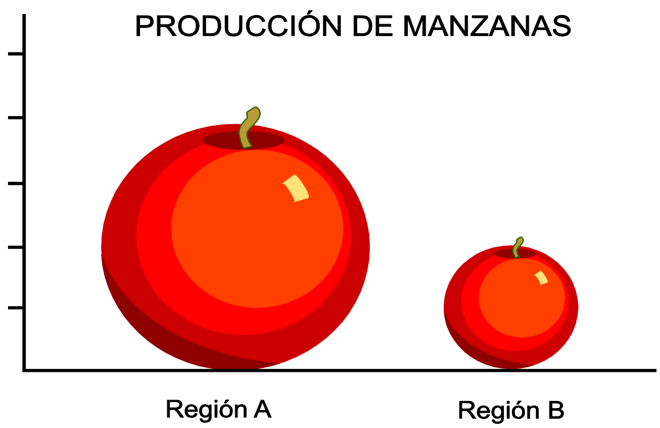
Pero si se quiere ser **precisos** en la información que se desea transmitir, debemos preferir los **gráficos estadísticos**.

Vamos a dar un ejemplo:

EJEMPLO 15

En un informe sobre la producción de manzanas, se pretendió mostrar un gráfico que permitiera comparar la producción de esta fruta de la Región A con la de la Región B.

Con el objeto de atraer la atención del lector, en lugar de un gráfico de barras se elaboró un bonito infograma, como el que se muestra a continuación:



Muy atractivo, pero no queda claro si las producciones son proporcionales a la **altura** de las manzanas, al **área visible** de las manzanas, o al **volumen** de ellas.

Seguramente sabes que el área de un círculo de diámetro d está dada por $\pi d^2/4$ y el volumen de una esfera por $\pi d^3/6$ (las manzanas se dibujaron con forma de esfera).

El gráfico muestra que el diámetro de la manzana grande es el doble de la correspondiente a la manzana pequeña.

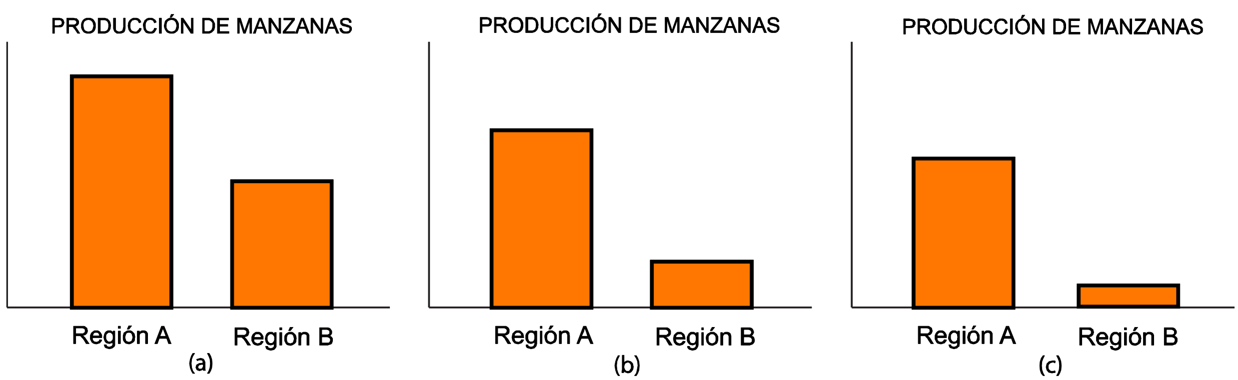
Tomemos el diámetro de la manzana grande como 1 y el de la pequeña como 0,5.

Las alturas, las áreas y los volúmenes por cada Región están dadas en la siguiente tabla.

La tabla también contiene la producción de la Región B como porcentaje de la producción en la Región A.

Interpretación	Región A	Región B	Porcentaje de B respecto de A
Altura	1	0,5	50 %
Área	π/4	0,25π/4	25 %
Volumen	π/6	0,125π/6	12,5 %

Las tres interpretaciones posibles del infograma se muestran en los siguientes gráficos de barras:



¿Qué tal? Confuso el infograma, ¿no es cierto?

EJERCICIOS

9) Con los datos de las detenciones de la envasadora de jugo, del Ejercicio 8, construye un gráfico de barras.

10) Los siguientes datos corresponden a los valores de precipitación de azúcar medido en la sangre de 63 individuos:

91	46	112	87	92	108	94	132	83
115	71	97	87	78	96	112	123	125
112	82	33	95	111	102	87	101	132
108	113	58	91	103	95	126	112	67
108	88	78	113	94	95	125	134	89
95	100	90	101	53	102	85	132	97
136	98	87	77	115	65	88	90	94

a) Construye una tabla de frecuencias por intervalos. Debes decidir cuántos intervalos y de qué largo.

b) A partir de la tabla, construye un histograma. ¿Observando el histograma, qué puedes decir?

Medidas de centro: la media y la mediana

Hay medidas que se **calculan** a partir de datos numéricos, y que sirven para **resumir** esos datos.

Son medidas que **resumen** la información contenida en los datos.

Se espera que estas medidas **describan** alguna característica del conjunto de datos.

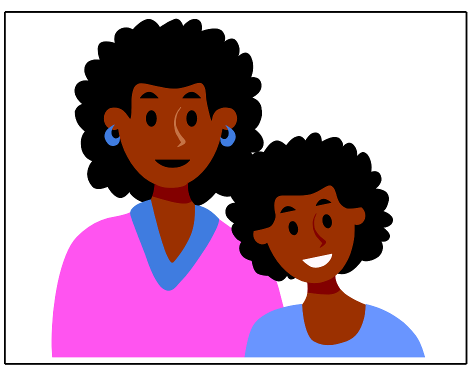
Unas medidas que describen un conjunto de datos, que seguramente ya conoces, son las **medidas de centro**.

Estas indican lo que se puede considerar como el **centro** de un conjunto de datos.

Hay más de una medida de centro, y las vamos a repasar a continuación.

La **media** o **promedio** es la suma de los valores dividido por el número de datos.

Entonces ocurre que no necesariamente el promedio se expresa en la misma escala que las observaciones originales.



Según el Instituto Nacional de Estadísticas, en 2017 el número promedio de hijos por cada mujer era de 1,6. Ninguna mujer tiene 1,6 hijos, pero como promedio es aceptable que no sea entero.

EJEMPLO 16

Supongamos que el profesor de Ciencias Sociales toma una prueba a un curso de 25 estudiantes.

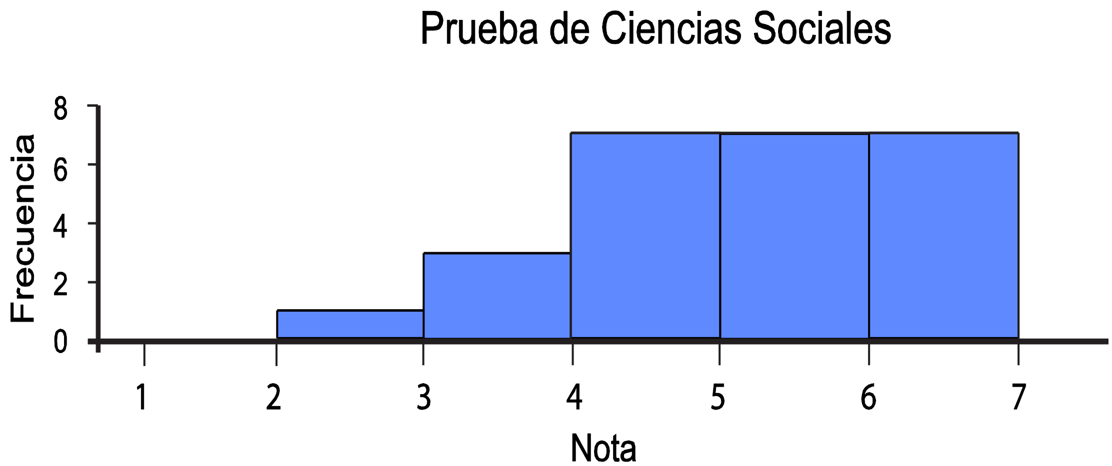
El profesor corrige la prueba y registra las notas.

Él quiere resumir los resultados, para tener una idea de cómo respondió el curso al aprendizaje de la materia que se evaluó en la prueba.

Una buena medida para resumir es obtener el promedio de las notas del curso.

Supongamos que las siguientes son las notas obtenidas por sus alumnos, ordenadas según el orden alfabético de los apellidos, que acompañamos por un histograma:

5,2	4,6	6,2	5,9	7,0
4,9	5,8	3,2	6,1	5,8
6,5	4,9	2,4	6,0	4,8
3,2	4,8	5,9	6,2	4,7
6,8	3,2	5,3	4,2	6,5

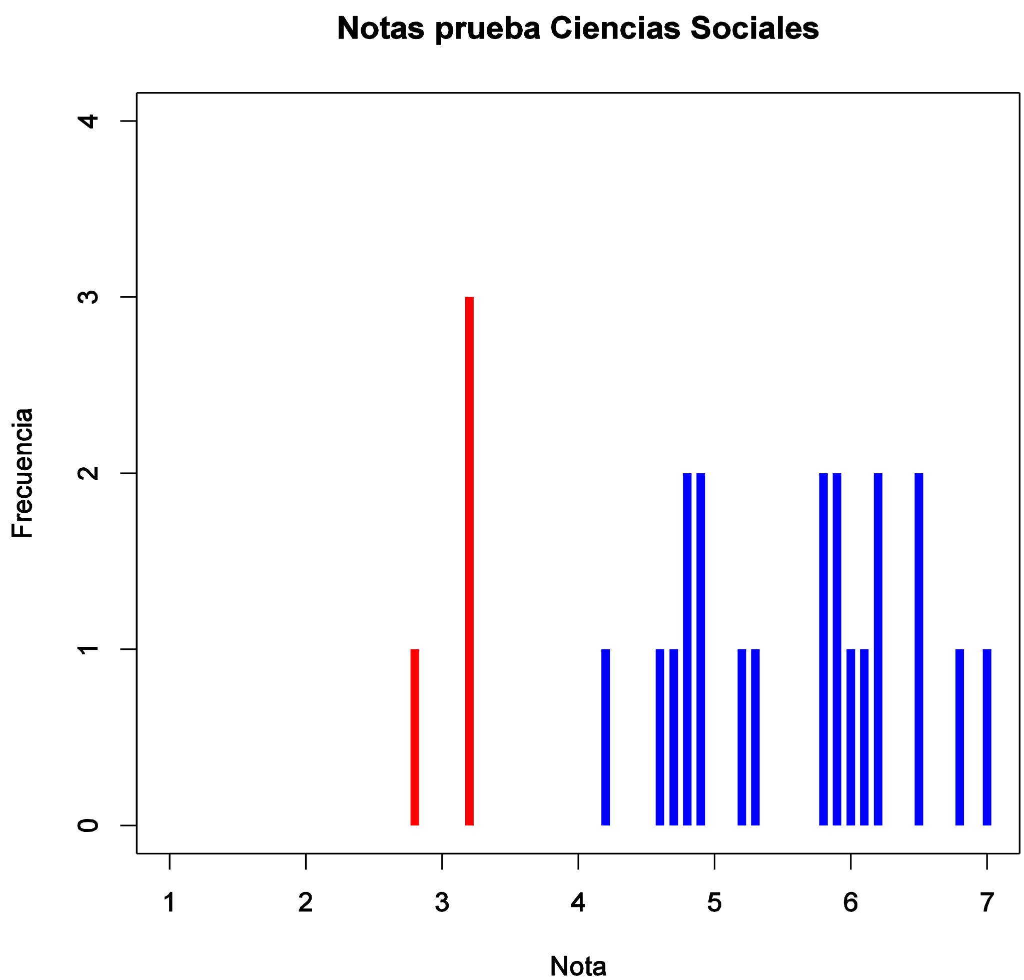


Puedes verificar que las notas suman 130,5 y si la dividimos por el total de alumnos, 25, obtenemos el promedio 5,22

Observa que las notas tienen una décima, pero el promedio está expresado con décimas y centésimas.

Se puede interpretar como que el **centro** de los datos es 5,22. O bien como si todo el curso completo obtuvo un 5,22.

El siguiente gráfico de barras ilustra el conjunto de las notas (en rojo las que están bajo 4,0).



Pero hay otras medidas de centro.

Una es la **mediana**. Es un número tal que al menos el 50% de los datos es menor o igual que él, y al menos el 50% de los datos es mayor o igual a él.

Es complicada la definición, pero en la práctica se obtiene de la siguiente manera, muy sencilla:

Se ordenan de menor a mayor.

Si hay un número impar de datos, la mediana es el que quedó al centro.

Si hay un número par de datos, es el promedio de los dos que quedaron al centro.

EJEMPLO 17

Continuación del Ejemplo 16, de las notas de la prueba de Ciencias Sociales. Calcularemos la mediana.

Primero viene la parte más difícil, ordenar las notas de menor a mayor. Quedan así:

2,4	3,2	3,2	3,2	4,2
4,6	4,7	4,8	4,8	4,9
4,9	5,2	5,3	5,8	5,8
5,9	5,9	6,0	6,1	6,2
6,2	6,5	6,5	6,8	7,0

Hay 25, luego la mediana es la nota del centro, la del lugar 13, que corresponde a un 5,3.

Podemos observar que es parecido al promedio, 5,22 pero no igual.

Es simplemente otra medida de centro.

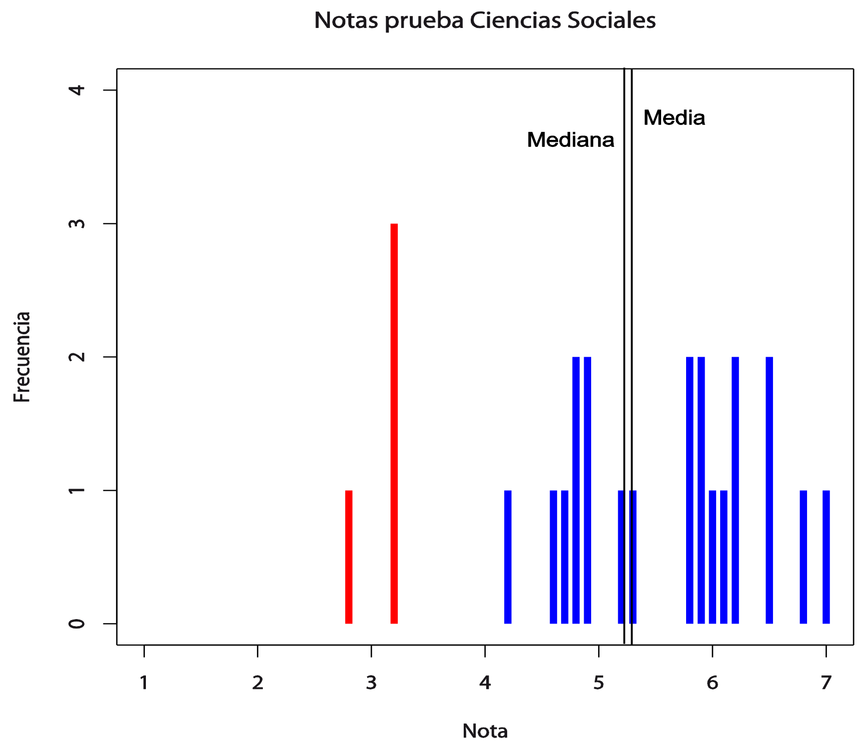
La figura siguiente muestra un gráfico de barras de las notas, con las medidas de centro.

Antes de hacer el gráfico, una observación:

Siempre que hagamos cualquier tipo de gráfico, no debemos olvidar de ponerle un título descriptivo y de ponerle títulos a los ejes.

Además, si corresponde, agregar la unidad de medida de las variables representadas en los ejes.

Un gráfico debe entenderse por sí solo. Si no, simplemente no sirve.



Hemos marcado la ubicación del promedio y de la mediana, para ver que son ambas medidas de centro, aunque difieren un poco entre ellas.

¿Cuál es la diferencia entre la media y la mediana?

La media es más fácil de calcular.

Intenta sumar estos 20 números, sin usar calculadora, y tómate el tiempo:

3, 1, 10, 12, 2, 8, 14, 4, 10, 5, 1, 14, 12, 8, 7, 4, 4, 9, 11, 15.

Ahora ordénalos de menor a mayor, y tómate el tiempo.

Con toda seguridad te demoraste más en ordenarlos que en sumarlos.

Otra diferencia muy importante entre la media y la mediana:

La media es muy **sensible** a **valores extremos**, mientras que la mediana es poco sensible.

Cuando los datos presentan **asimetría**, es decir, contienen valores muy alejados del centro, o muy grandes o muy pequeños, es mejor usar la mediana, como medida de centro, pues la media aparece distorsionada por los valores extremos.

Otra medida de centro, aunque no muy buena, es la moda.

Consiste en el valor que más se repite.

En el ejemplo de las notas la moda es el 3,2. Te darás cuenta de que no está muy al centro.

En general, la moda sirve como medida de centro cuando hay una cantidad muy grande de datos.

Hay conjuntos de datos que tienen más de una moda.

Medidas de posición: los Percentiles

Se puede generalizar la idea de la mediana y definir otras medidas que no corresponden a un centro, sino que a otras posiciones dentro del conjunto de datos.

Estas medidas son los percentiles, que dividen el conjunto de datos en dos grupos con determinados porcentajes de observaciones.

Primero daremos una definición y luego presentaremos un método para encontrar un percentil.

Sea q un número entero entre 1 y 99. El percentil q, que designaremos P_q, es un número tal que al menos $q\%$ de los datos son menores o iguales a él, y al menos $(100 - q)\%$ de los datos son mayores o iguales a él.

La mediana, entonces, es un caso especial de percentil. Es el percentil 50, puesto que el 50% de los datos son menores o guales a ella y el 50% de los datos son mayores o iguales.

Como en el caso de la mediana, la definición anterior es un tanto difícil de entender, pero hay una forma simple de obtener los percentiles, que es la siguiente:

Se quiere el percentil q, con q un entero entre 1 y 99.

Primero hay que ordenar los datos de menor a mayor, donde esta es la parte más complicada.

Después se debe calcular

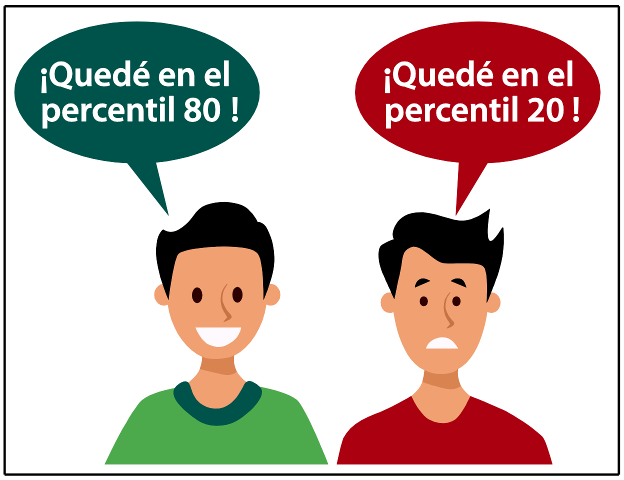
$$R = q \cdot \frac{n}{100}$$

Si R es entero, el percentil q es el promedio de la observación que ocupa el lugar R y la que ocupa el lugar $R+1$ en el conjunto ordenado. Si R no es entero, se debe aproximar al entero superior. Sea E ese entero. El percentil q es la observación que ocupa la posición E en el conjunto ordenado de datos.

EJEMPLO 18

Continuaremos con el Ejemplo 17 de las notas de Ciencias Sociales.

Ahora supongamos que el profesor de Ciencias Sociales desea saber cuáles son los percentiles 25 y 7.



Las notas ya las tenemos ordenadas en el Ejemplo 17. Recordemos que son 25 notas.

Primero veamos el percentil 25:

$$R = 25 \cdot \frac{25}{100} = 6{,}25$$

No es entero, luego lo aproximamos al entero superior $E = 7$.

El percentil 25 es la nota que ocupa el séptimo lugar, es decir, la nota 4,7.

Significa que al menos el 25% de las notas son iguales o inferiores a 4,7 y al menos el 75% de las notas son iguales o superiores a 4,7.

Efectivamente, si contamos, vemos que 7 notas son menores o iguales a 4,7, eso representa el $100 \cdot \frac{7}{25} = 28\%$ de las observaciones.

Y podemos ver que 19 notas son mayores o iguales a 4,7.

Esto representa un $100 \cdot \frac{19}{25} = 76\%$ de las observaciones.

El 4,7 cumple con la definición de percentil 25.

Ahora el percentil 75:

$$R = 75 \cdot \frac{25}{100} = 18{,}75$$

Nuevamente tenemos un resultado no entero, luego aproximamos R al entero superior, y obtenemos E=19.

El percentil 75 es la nota que ocupa el lugar 19, vale decir, el 6,1.

Entonces, el 75% de las notas son menores o iguales a 6,1 y el 25% son mayores o iguales a 6,1.

Veamos si efectivamente es el percentil 75.

Resulta que 19 notas son menores o iguales a 6,1, eso representa el $100 \cdot \frac{19}{25} = 76\%$ de las observaciones.

Y 6 notas son mayores o iguales a 6,1, lo que representa un $100 \cdot \frac{7}{25} = 28\%$ de las observaciones.

La nota 6,1 cumple con la definición de percentil 75.

Ahora podríamos calcular el percentil 50, que debería ser igual a la mediana.

$$R = 50 \cdot \frac{25}{100} = 12{,}5$$

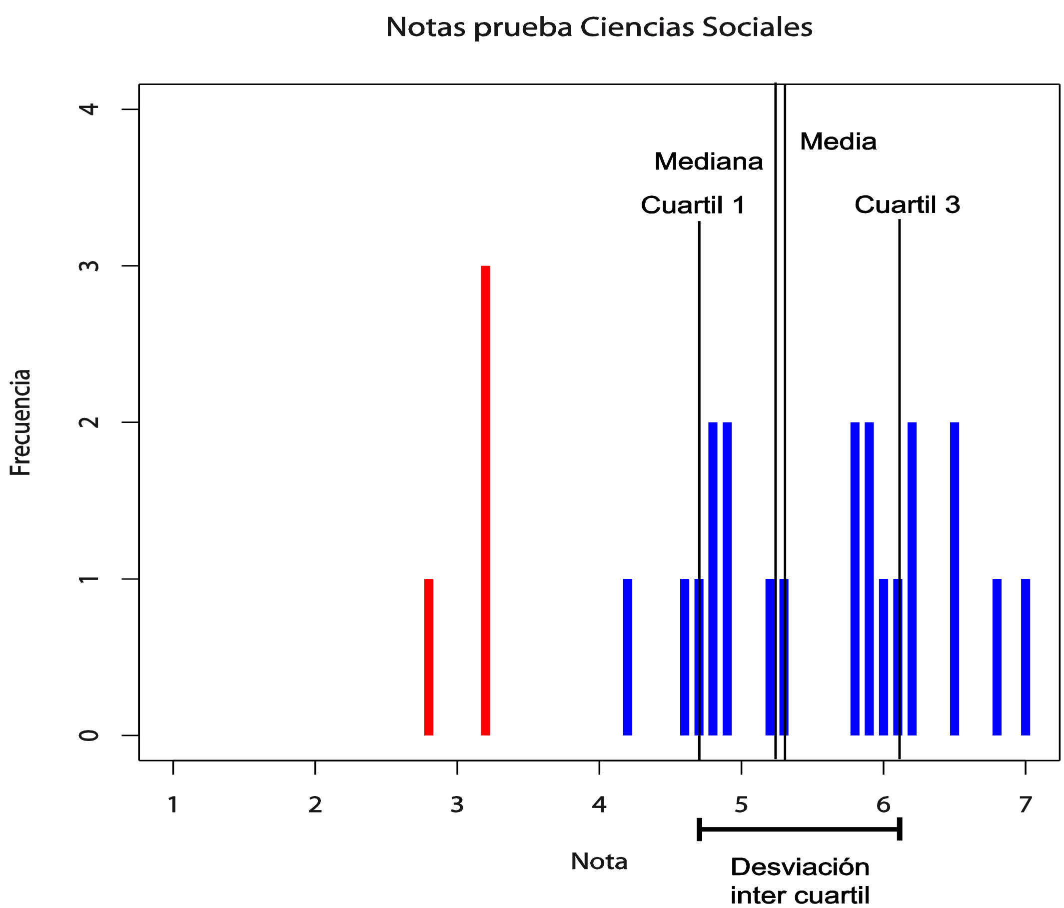
Se debe aproximar al entero superior. Nos da la posición 13.

Busquemos la nota que está en el lugar 13 del conjunto ordenado, y corresponde a la mediana 5,3 que ya habíamos calculado en el Ejemplo 17.

La figura siguiente muestra el mismo gráfico de barra que el Ejemplo 17, pero se han agregado dos líneas que muestran la posición de los percentiles 25 y 75, también llamados **cuartil 1** y **cuartil 3**, respectivamente, porque dividen el conjunto de datos en cuatro.

La mediana es también el **cuartil 2**.

La distancia entre los dos cuartiles se denomina **desviación intercuartil**, y está representada en el gráfico. Es una medida de cuán dispersos están los datos.



Dijimos que los percentiles 25, 50 y 75, también se llaman **cuartiles**. Se simbolizan por Q_1, Q_2 y Q_3.

De forma similar, si consideramos los percentiles 20, 40, 60 y 80, también se denominan **quintiles**, porque dividen el conjunto de datos en cinco grupos, con un 20% en cada uno. Son los quintiles 1, 2, 3 y 4.

Los otros que tienen un nombre particular son los percentiles 10, 20, 30, 40, 50, 60, 70, 80, y 90, que dividen el conjunto de datos en 10 grupos con un 10% en cada grupo.

Obviamente, estos se llaman **deciles**. Son los deciles 1 a 10.

Algunas de **observaciones**:

Recordemos que dijimos que la **mediana** es una medida de centro **robusta**.

Eso significa que, a diferencia de la media, está muy poco afectada por observaciones extremas.

Pues bien, los **percentiles** son todos **medidas de posición robustas**. Las observaciones extremas los afectan muy poco.

Otra **observación**: es posible que haya más de un número que cumple con la definición que dimos de **percentil**.

Por eso hay más de un método de cálculo de percentiles. Pero todos llegan a valores muy parecidos.

Tercera **observación**: en jerga popular suelen referirse a los quintiles como cada **grupo** de 20% de datos contenido entre dos quintiles contiguos.

Eso es un error: los quintiles son los **números que dividen** el conjunto de datos en grupos de 20% en cada uno, y no el grupo mismo.

EJERCICIOS

11) Con los datos de las detenciones de la máquina envasadora de jugo, del Ejercicio 8, obtén la media, la mediana y los quintiles.

No te olvides de las unidades de medida.

12) Con los datos del Ejercicio 10, de la precipitación de azúcar en la sangre, obtén la media, la mediana y los cuartiles.

El diagrama de cajón con bigotes

También llamado diagrama de **cajón** o de **caja**, o bien **cajagrama**.

Es un gráfico construido en base de medidas **robustas**.

Por eso tiene la ventaja de estar poco influenciado por datos extremos.

Es más, nos permite visualizar con claridad cuáles datos son efectivamente extremos, si los hay.

Para construir un gráfico de cajón se traza una línea auxiliar con una escala numérica. Lo normal es que sea horizontal, pero también puede ser vertical.

Luego se dibuja un rectángulo, cuyos extremos se posicionan en los cuartiles Q1 y Q3.

Dentro del rectángulo se dibuja una línea posicionada en la mediana.

La distancia $RIC = Q3 - Q1$ es el **rango intercuartil.**

Medimos la distancia entre el lado izquierdo del rectángulo hasta 1,5 veces el RIC, hacia la izquierda y hacemos una marca.

También medimos la distancia entre el lado derecho del rectángulo hasta 1,5 veces el RIC hacia la derecha, y hacemos otra marca.

Todos los datos que están entre las dos marcas se llaman **datos interiores** y se considera que no son extremos.

Si hay datos que estén fuera de las dos marcas, hacia la izquierda o hacia la derecha, se consideran **valores extremos**.

Los **bigotes** son trazos que van desde los lados del rectángulo hasta los valores mínimo y máximo, respectivamente, de los datos interiores.

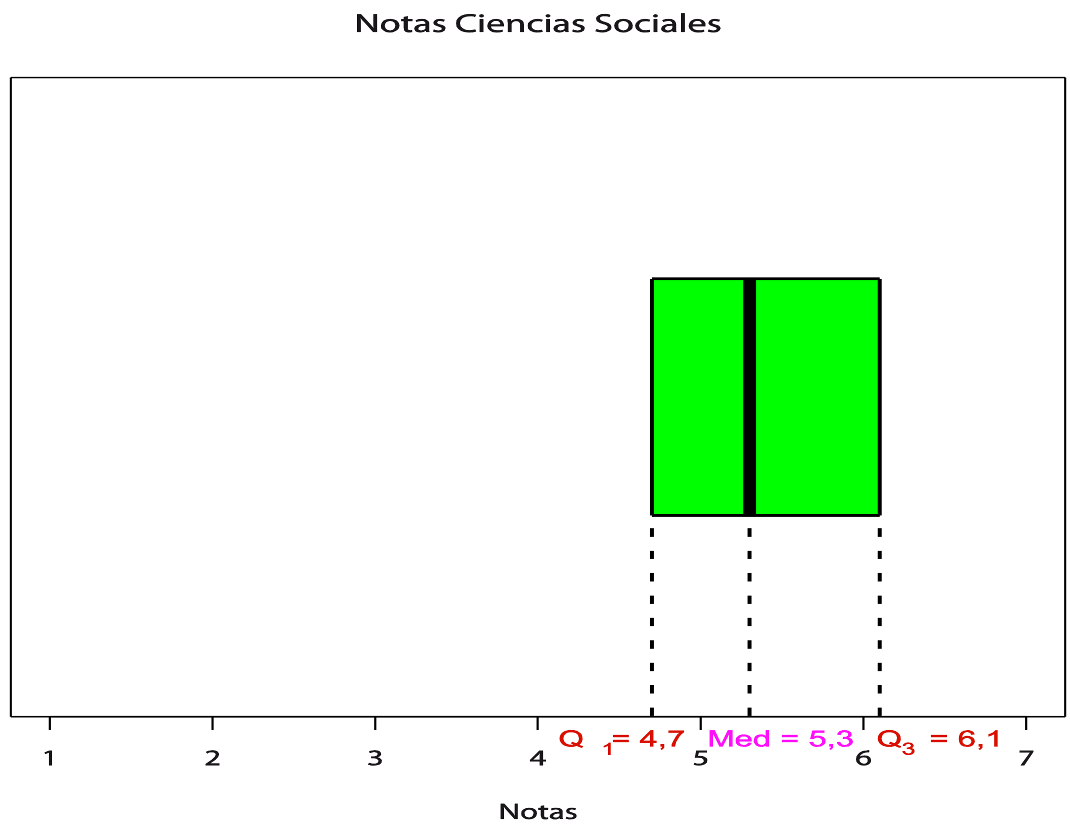
Si hay valores extremos, es recomendable estudiarlos con el objeto de determinar por qué son extremos, si realmente se escapan del grupo de datos, o hay un error, ya sea de observación, de tipeo o de cálculo.

EJEMPLO 19

Vamos a construir un cajón con bigotes con las notas de la prueba de Ciencias Sociales, del Ejemplo 16.

En los Ejemplos 17 y 18 obtuvimos la mediana y los cuartiles (percentiles 25 y 75), así que gran parte del trabajo está hecho. Vimos que la mediana es 5,3 y los cuartiles son 4,7 y 6,1.

El cajón tiene por lados los cuartiles, y contiene un trazo vertical que representa la mediana. El cajón se está viendo como muestra la figura siguiente:



Ahora tenemos que agregarle los bigotes. El **Rango intercuartil** es

6,1 - 4,7 = 1,4.

Hay que multiplicarlo por 1,5, lo que da 2,1.

Si se le resta 2,1 al cuartil 1 y se le suma 2,1 al cuartil 3, obtenemos 2,6 y 8,2 respectivamente.

La menor de las notas mayores o iguales a 2,8 y la mayor de las notas menores o iguales a 8,2 son los extremos de los **bigotes**.

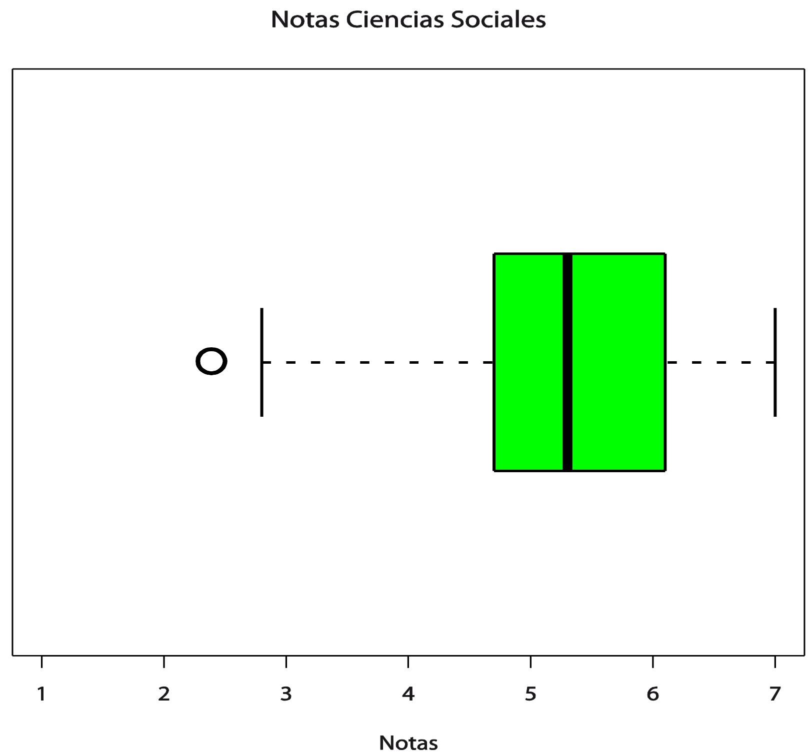
Ellos son, respectivamente, las notas 3,2 y 7,0.

Cualquier nota menor que 2,8 o mayor que 8,2 es considerada una observación extrema.

Es obvio que no hay ninguna mayor que 8,2. Pero sí hay una menos que 2,6, el 2,4 que obtuvo un alumno desafortunado.

Tales datos se suelen representar por círculos o asteriscos.

El siguiente es el cajón con bigotes terminado, al que no hemos olvidado ponerle título:



Ahora veremos otro uso de las medidas de centro. Se trata de evaluar la **asimetría** que pueden presentar los datos.

La asimetría se forma por la presencia de datos extremos, hacia uno de los lados, ya sea que sean más grandes que la mayoría de los datos restantes, o ya sea que son más pequeños que la mayoría de los datos.

Dijimos que la media o promedio está muy influenciada por **observaciones extremas**, mientras que la mediana no lo está.

Entonces la comparando la media con la mediana tendremos una idea de la asimetría.

Si la media es más grande que la mediana, hay asimetría en forma de una cola hacia el lado de los valores grandes.

Si la media es más pequeña que la mediana, la asimetría tiene forma de una cola hacia el lado de los valores pequeños.

Y si la media y la mediana son parecidas, quiere decir que los datos son aproximadamente **simétricos**.

En los siguientes tres ejemplos veremos diferentes situaciones:

EJEMPLO 20

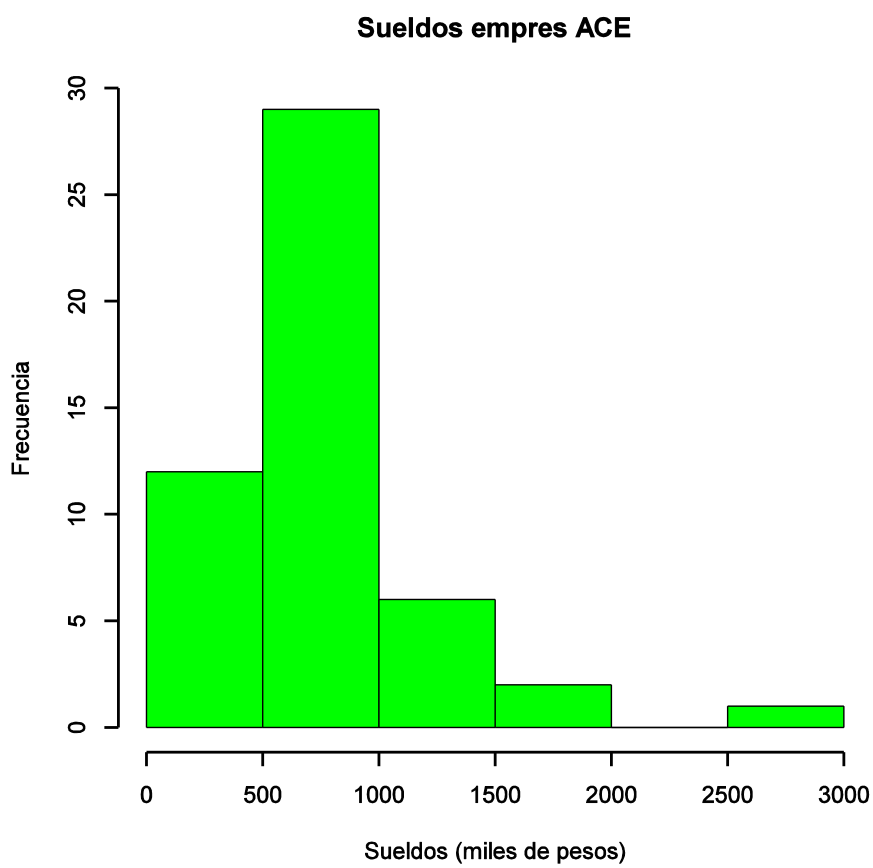
El primer conjunto de datos corresponde a los sueldos de los 50 empleados de la empresa ACE, en miles de pesos.

Los datos se muestran ordenados de menor a mayor, para facilitar los cálculos que llevaremos a cabo.

320	326	326	345	347	370	379	394	402	409
415	505	527	581,2	585	587	588	622	639	660
671	678	689	694,5	707	746	746	776	790	741
767	787	806	821,2	850	888	940	955	986	996
1033	1128	1165	1171	1181	285	1455	1682	1887	2675

Veremos un histograma de estos datos, agrupados en intervalos de 500 mil pesos. No hemos olvidado ponerle título al gráfico y a los ejes.

El eje horizontal, de los sueldos, debe llevar la unidad de medida (miles de pesos):



Observa que claramente hay un grupo grande de empleados cuyos sueldos están concentrados en valores menores que un millón, pero hay algunos que se escapan hacia valores mayores.

Esto es una **asimetría** en los datos, Particularmente, es una **asimetría positiva** o **a la derecha**.

En datos reales, es muy frecuente que se presenten dos características: sólo valores positivos y asimetría a la derecha.

Esto es frecuente en datos como pesos, tiempos, longitudes, capacidades, etc.

Veamos cómo son las medidas de centro y los cuartiles, en este conjunto de datos:

Cuartil 1: 510,5 mil pesos

Mediana: 700.8 mil pesos

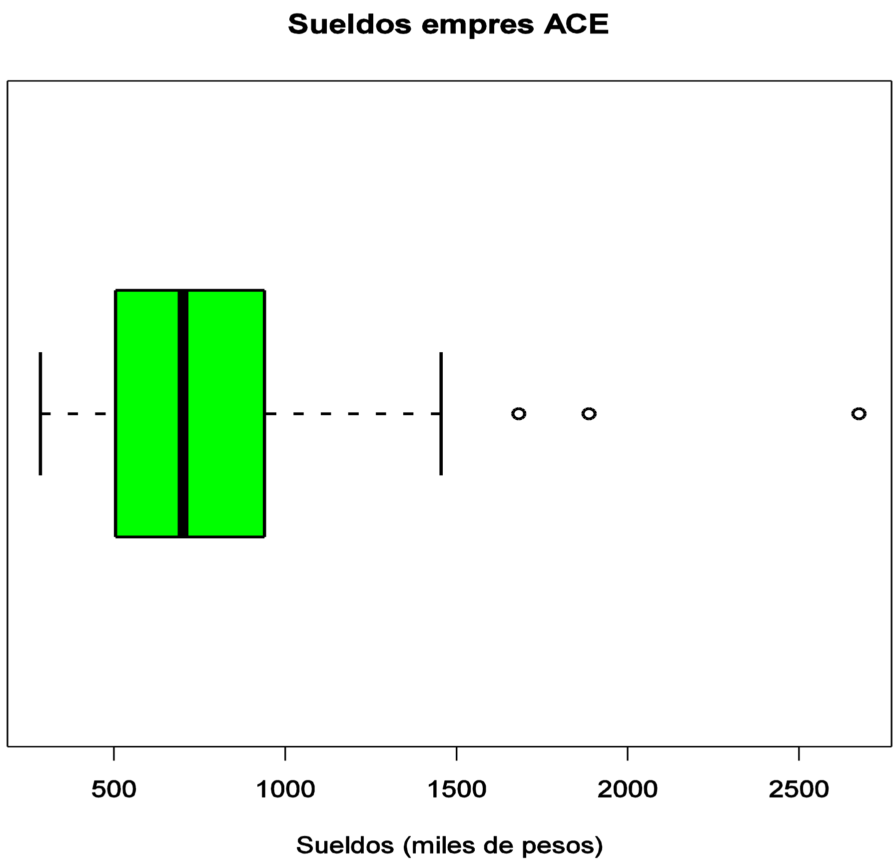
Promedio o media: 780.8 mil pesos

Cuartil 3: 927.0 mil pesos.

Observa que el promedio es más grande que la mediana, aun cuando ambas son medidas de centro.

Por supuesto que tiene que ser así, pues recuerda que el promedio está afectado por valores extremos, en cambio la mediana no: es una medida **robusta**.

Con estas medidas podemos construir el cajón con bigotes, que se muestra a continuación:



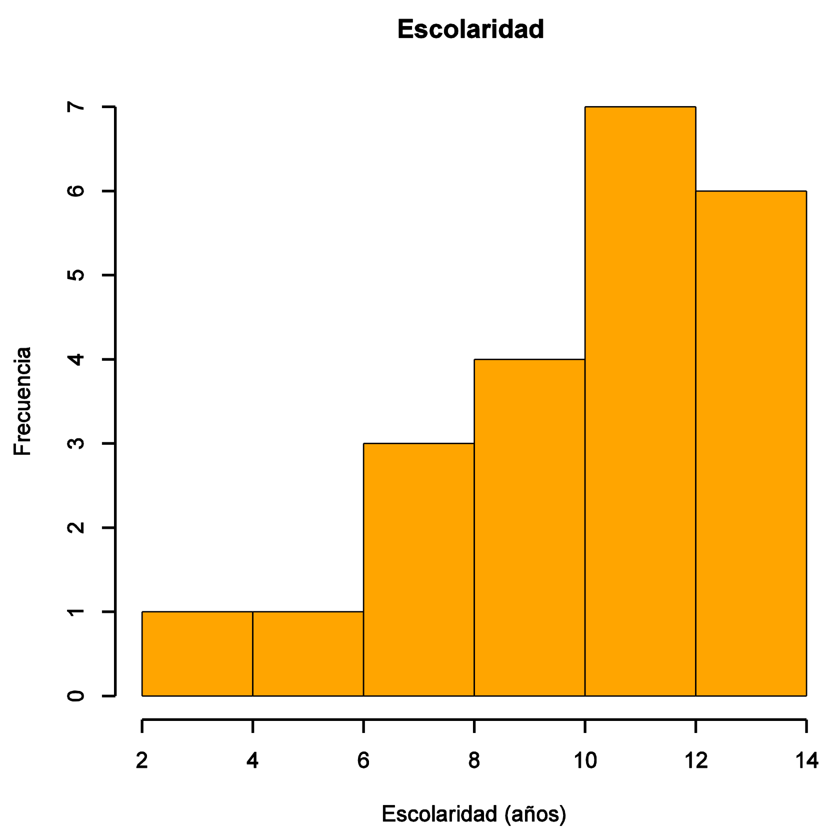
En este gráfico se aprecia la asimetría positiva o derecha, y quedan individualizados tres valores extremos, que corresponden a los sueldos más altos, 1682, 1887, y 2675, en miles de pesos.

EJEMPLO 21

El segundo conjunto corresponde a los años de escolaridad de una muestra de 22 habitantes de un pueblo rural.

3	5	8	8	8	10	10	10	10	11	11
12	12	12	12	12	13	14	14	14	14	14

Los datos se resumen en el siguiente histograma, con intervalos de 2 años:



En este caso hay una asimetría inversa a los datos de los sueldos: ahora la cola está hacia la izquierda.

Se dice que la asimetría es **negativa** o **a la izquierda**.

Ahora las medidas de centro y los cuartiles son los siguientes:

Cuartil 1: 22 años

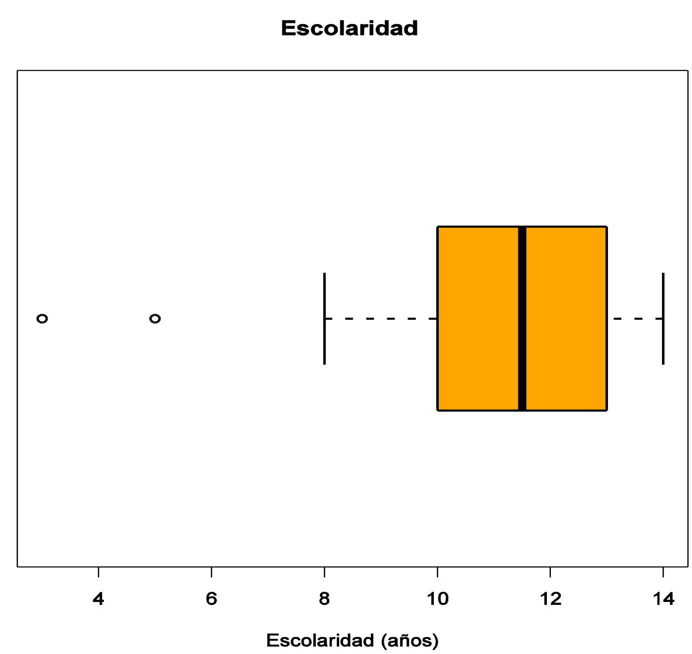
Mediana: 11,5 años

Promedio o media: 10 años

Cuartil 3: 12,75 años.

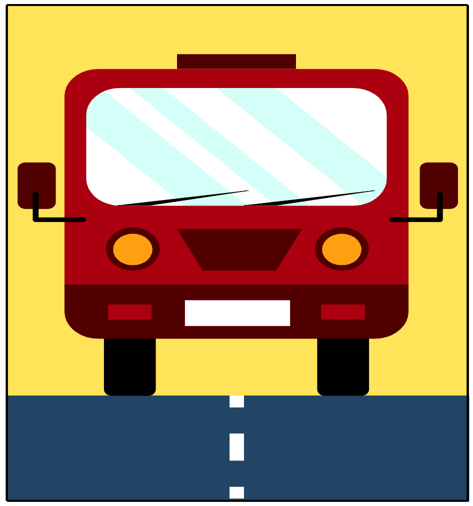
En este caso, al contrario de los sueldos, el promedio es más pequeño que la mediana, pues el promedio está influido por los datos extremos que son pequeños, y la mediana no.

Construimos el gráfico de cajón con estas medidas, y observamos dos valores extremos:



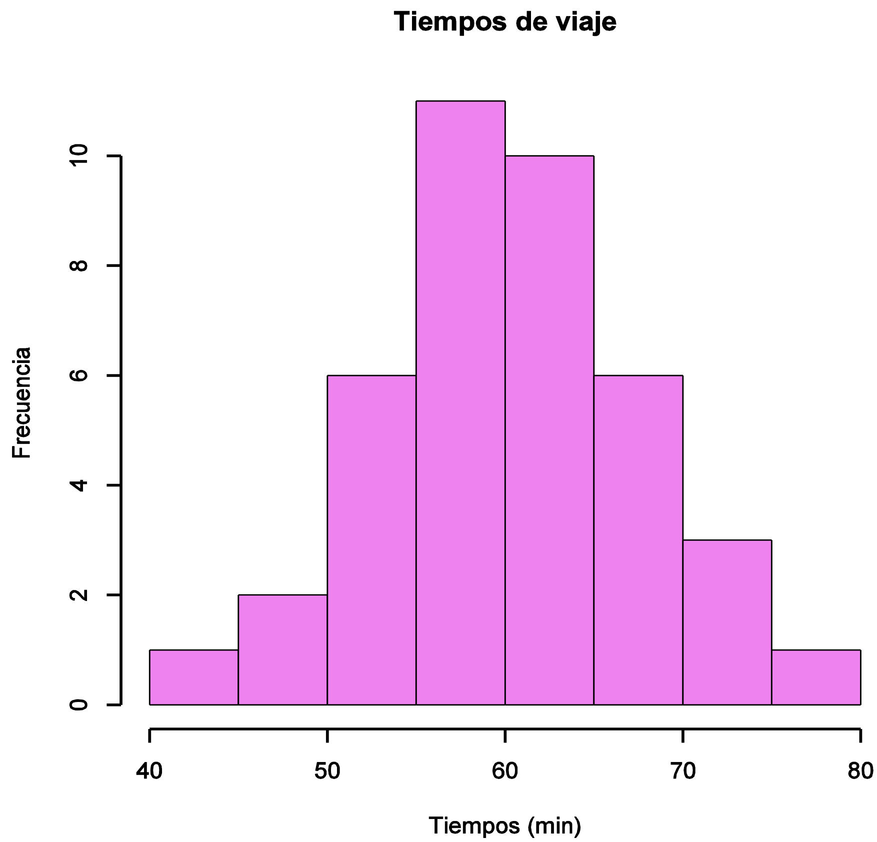
EJEMPLO 22

El tercer conjunto muestra los tiempos de viaje en locomoción colectiva de una muestra de 40 trabajadores, en minutos.



44	48	49	51	53	53	55	55	55	56
56	58	58	59	59	59	59	60	60	60
61	61	61	61	61	61	61	63	63	65
66	67	67	67	69	70	72	74	75	76

El histograma, con intervalos de 5 minutos, se muestra a continuación:



Ahora vemos que, a diferencia de los anteriores conjuntos de datos, es bastante simétrico.

Las medidas de centro y los cuartiles son los siguientes:

Cuartil 1: 56 minutos

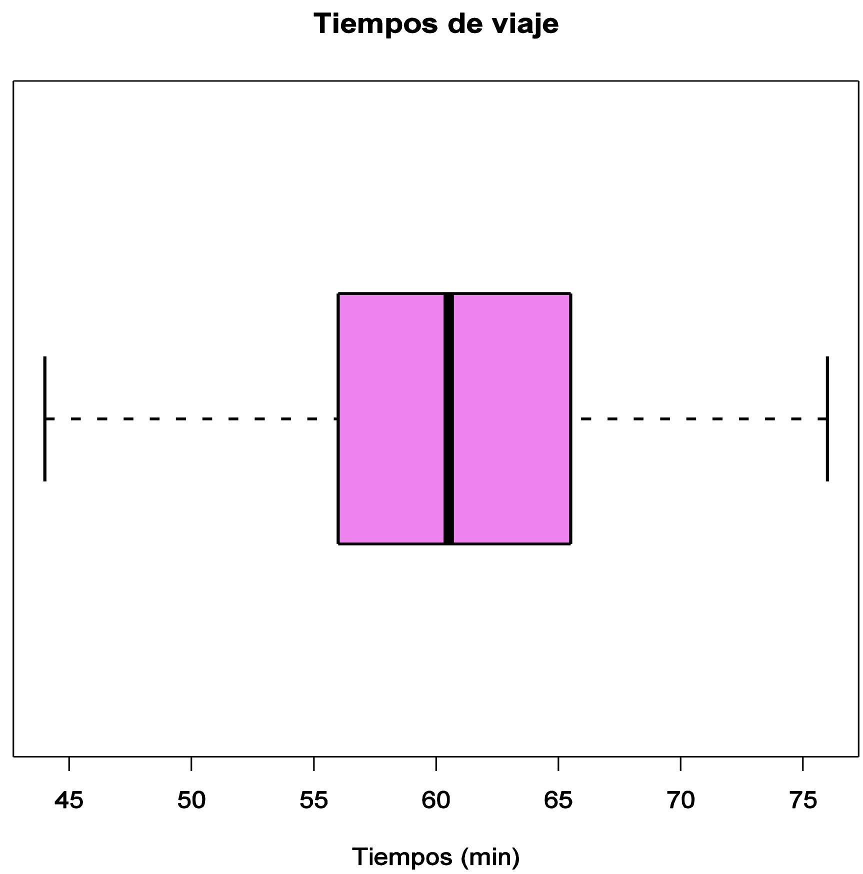
Mediana: 60,5 minutos

Promedio o media: 60,7 minutos

Cuartil 3: 65,25 minutos.

En este caso la media y la mediana casi coinciden, eso es consecuencia de que los datos son **simétricos**.

El gráfico de cajón es el siguiente:



También muestra la simetría. Y no presenta valores extremos en ninguno de los dos lados.

Podemos concluir que la diferencia de la media menos la mediana nos da una idea de la asimetría que presenta el conjunto de datos:

Si es positiva, hay asimetría a la derecha o asimetría positiva;

si es negativa, hay simetría a la izquierda o asimetría negativa,

si la diferencia es cercana a cero, hay simetría.

EJEMPLO 23

Vamos a hacer el siguiente experimento para comprobar la robustez o falta de ella de las medidas de centro y los cuartiles.

Tomemos el tercer conjunto de datos del Ejemplo anterior: los tiempos de viaje en locomoción colectiva de una muestra de 40 trabajadores, en minutos.

Recordemos que:

el Cuartil 1 es 56,0 minutos

La Mediana es 60,5 minutos

el Promedio es 60,7 minutos

el Cuartil 3 es 65,25 minutos.

Ahora vamos a suponer que, en lugar de 40 datos, hay 42, y que los que faltaron viven lejos de su lugar de trabajo, y se demoran 190 y 195 minutos, respectivamente.

Claramente son valores extremos. Si recalculamos las medidas, obtenemos los siguientes valores:

el nuevo cuartil 1 es 56,5 minutos

La mediana queda en 61,0 minutos

el promedio queda en 66,98 minutos

el cuartil 3 queda en 66,75 minutos.

Puedes observar que el promedio aumentó en 6,28 minutos, lo que, en porcentaje, corresponde a 100x6,28/60,7=10,3%. Bastante grande el cambio.

Por otro lado, la mediana aumentó en 0,5 minutos, lo que representa apenas un 100\times 0,5/60,7=0,82%.

El cambio en la mediana es mínimo por la introducción de estas dos observaciones extremas, en cambio el promedio aumentó más de un 10%.

Por eso decimos que la mediana es robusta, mientras que el promedio no lo es.

Si comparamos los cuartiles, vemos que el cuartil 1 y el cuartil 3 aumentaron en 0,5 y 1,5 minutos, respectivamente. En términos porcentuales son, respectivamente un 0,89% y un 0,77%.

Estos cambios son muy pequeños, comparados con el aumento en el promedio.

EJERCICIOS

13) En relación con el Ejemplo 20, realiza los cálculos para obtener el promedio, la mediana y los cuartiles. Construye el diagrama de cajón.

14) En relación con el Ejemplo 21, realiza los cálculos para obtener el promedio, la mediana y los cuartiles. Construye el diagrama de cajón.

15) En relación con el Ejemplo 22, realiza los cálculos para obtener el promedio, la mediana y los cuartiles. Construye el diagrama de cajón.

16) En relación con el Ejemplo 23, realiza los cálculos para obtener el promedio, la mediana y los cuartiles. Construye el diagrama de cajón.

17) En relación con los datos de las detenciones de la máquina envasadora de jugo, del Ejercicio 8, construye un diagrama de cajón.

Compáralo con el histograma.

18) En relación con los datos del Ejercicio 10, de la precipitación de azúcar en la sangre, construye un diagrama de cajón.

Compáralo con el histograma.

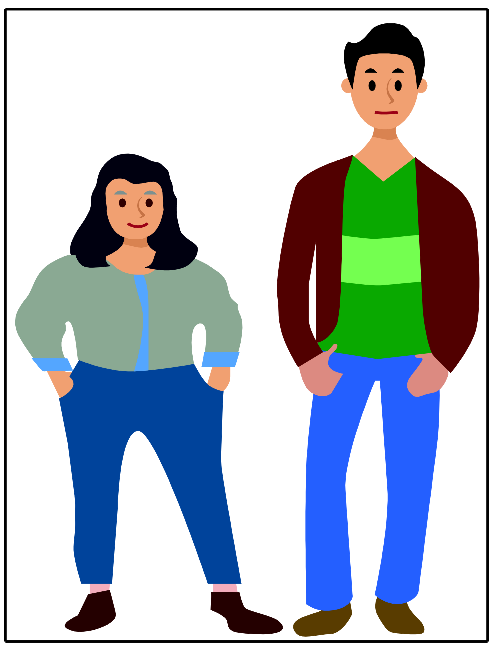
Asociación entre variables

MOTIVACIÓN

Consideremos las dos características de un grupo de personas: la altura y el peso.

Es evidente que en la mayoría de los casos las personas de mayor altura tienden a tener más peso.

En este caso decimos que hay una **asociación positiva** entre esas dos características.

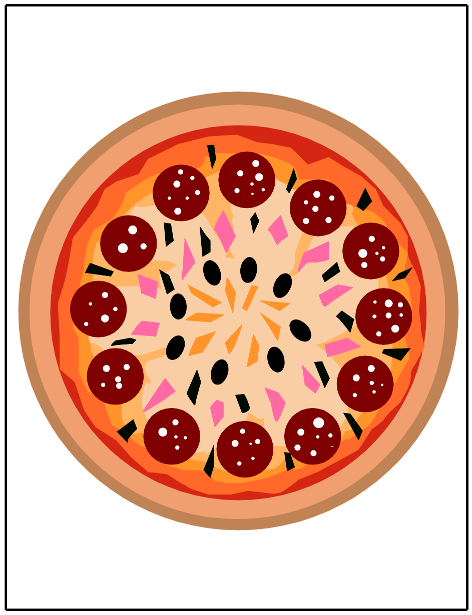


Aunque en algunos casos particulares puede haber una persona con más altura que otra, pero con menos peso, pero serían excepcionales.

En otra situación consideremos el precio de un producto, como, por ejemplo, las paltas, y el consumo de ese producto.

Si el precio de las paltas empieza a subir, las personas comienzan a consumir menos paltas, sustituyéndolas por otros productos. El consumo comienza a bajar.

Aquí decimos que las características **precio** y **consumo** de un producto tienen una **asociación negativa**.



Por último, consideremos las características **venta semanal de pizzas en Iquique** y **venta semanal de pizzas en Rancagua**.

Es muy posible que el aumento de una de ellas no implica un aumento de la otra.

Si es así, decimos que **no hay asociación** entre las características.

Consideremos dos variables medidas a un conjunto de objetos.

Veremos el grado en que estas dos variables están **asociadas**.

El que estén asociadas significa que, si una aumenta, la otra tiende a aumentar, o bien, si una aumenta, la otra tiende a disminuir.

Por el contrario, si una aumenta la otra puede aumentar, disminuir o mantenerse estable, sin un patrón definido, decimos que **no hay asociación** entre esas variables.

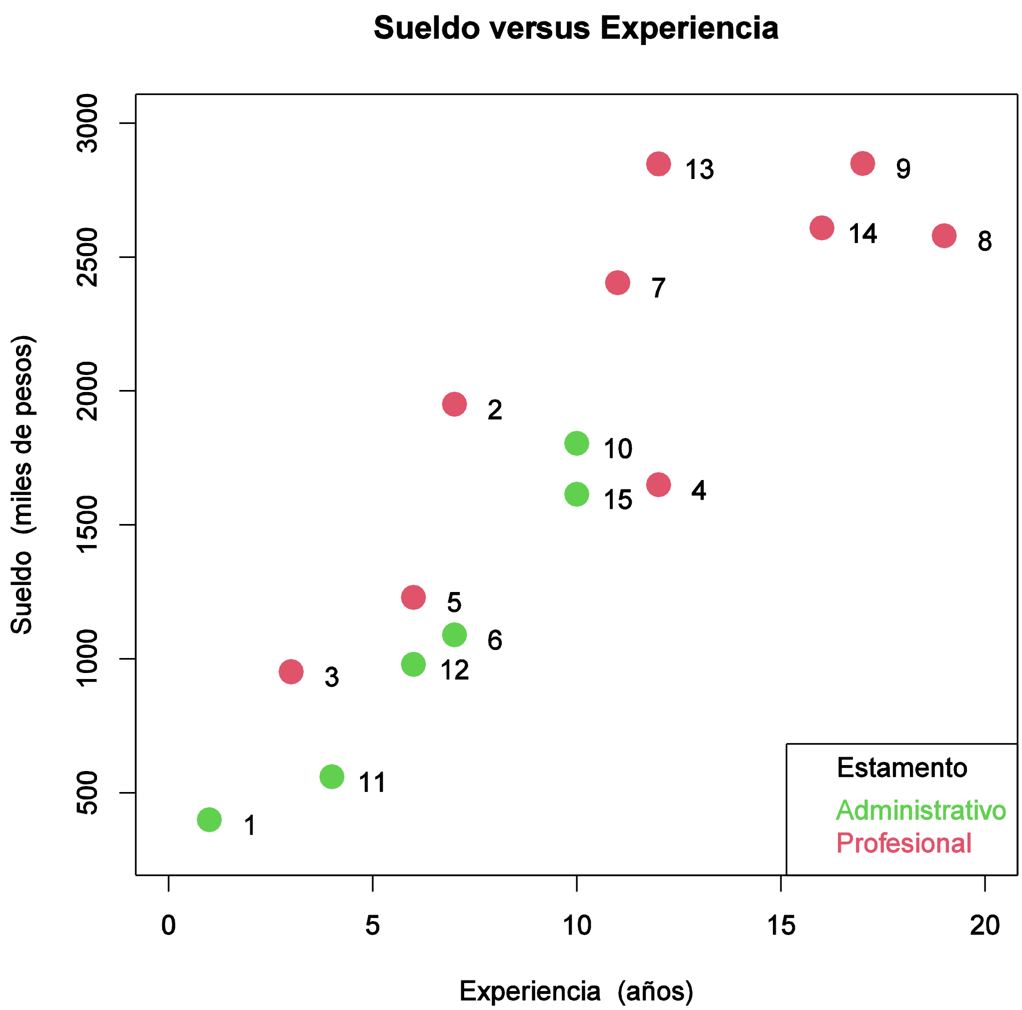
EJEMPLO 24

La tabla siguiente presenta lo años de experiencia y los sueldos de una muestra de 15 funcionarios profesionales.

El Estamento indica si es Profesional (1) o Administrativo (2).

Funcionario	X - Años de experiencia	Y - Sueldo en miles de pesos	Estamento
1	1	400	2
2	7	1951	1
3	3	952	1
4	12	1650	1
5	6	1230	1
6	7	1090	2
7	11	2405	1
8	19	2580	1
9	17	2850	1
10	10	1805	2
11	4	560	2
12	6	980	2
13	12	2848	1
14	16	2609	1
15	10	1615	2

Representaremos los datos en un gráfico de dispersión:



Se puede apreciar la asociación de tipo lineal existente entre ambas variables observadas.

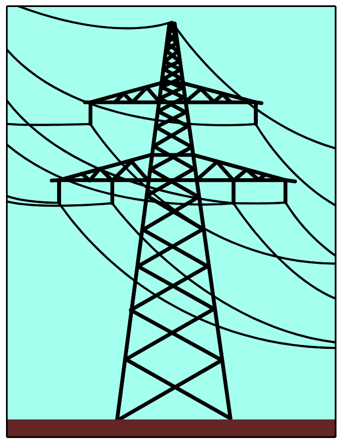
Es decir, los puntos parecen estar repartidos aproximadamente en torno a una línea diagonal imaginaria.

Además, esta línea tiende a subir a mediada que nos movemos hacia la derecha. O sea, si una variable crece, la otra tiende a crecer.

Como dijimos antes, esta es una **asociación positiva**.

EJEMPLO 25

Los datos siguientes corresponden a la generación de energía eléctrica durante el año 2007.



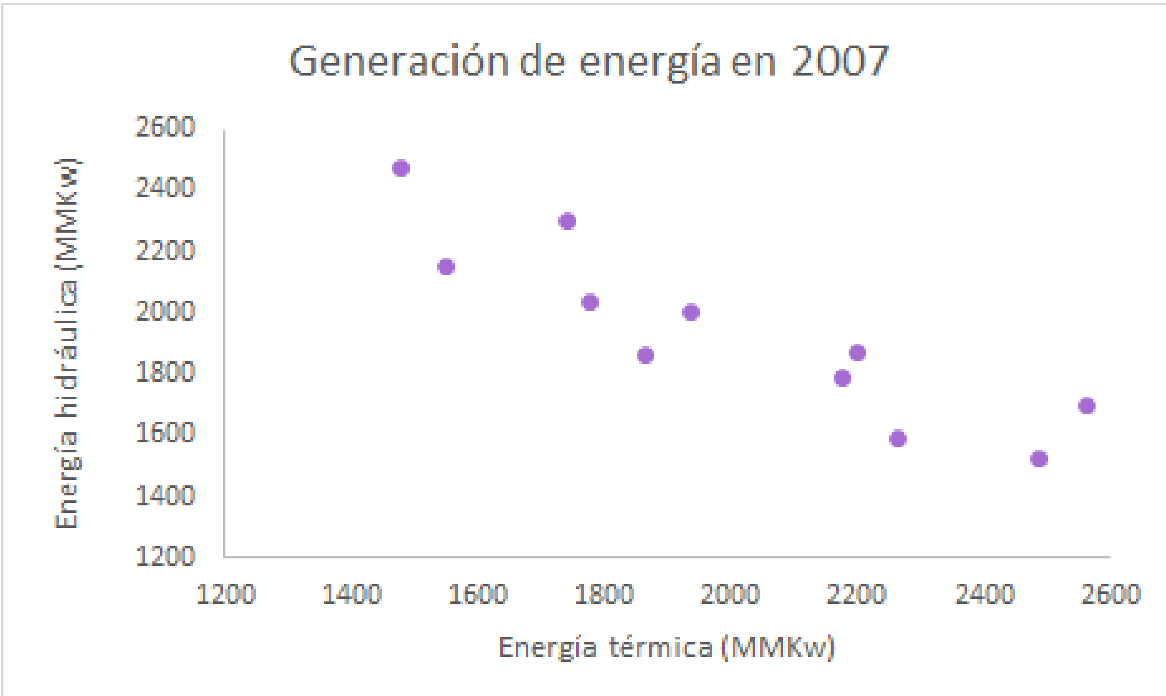
Mes	Trimestre	Térmica (X)	Hidráulica (Y)
Ene	1	1475	2480
Feb	1	1549	2158
Mar	1	1741	2307
Abr	2	1863	1868
May	2	2172	1790
Jun	2	2486	1525
Jul	3	2560	1700
Ago	3	2654	1427
Sep	3	2260	1594
Oct	4	1934	2004
Nov	4	1775	2036
Dic	4	2196	1874

Muestra el mes, el trimestre, los millones de kilowatts (MMKw) producidos por generación térmica (a partir de combustibles fósiles, como carbón, petróleo) y los millones de kilowatts producidos por generación hidráulica (por movimiento de agua debido a desniveles, que mueven turbinas).

Fuente: Banco Central de Chile

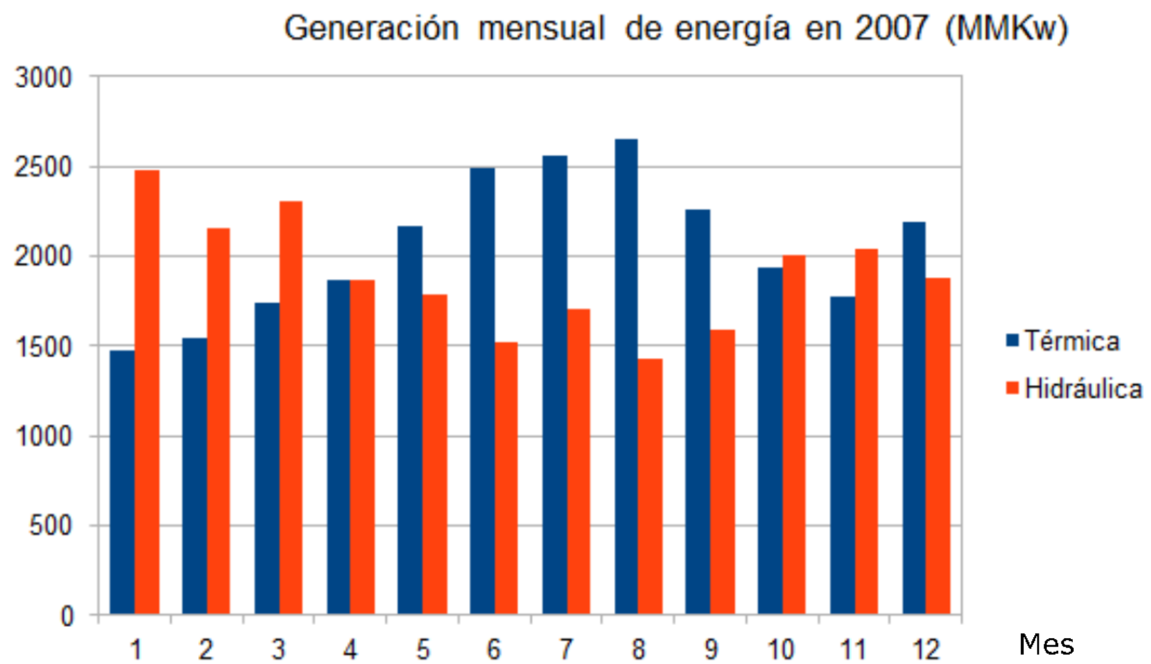
Primero usaremos Excel para visualizar estos datos.

El siguiente gráfico es un diagrama de dispersión que muestra la energía térmica generada versus la energía hidráulica, en cada uno de los meses.

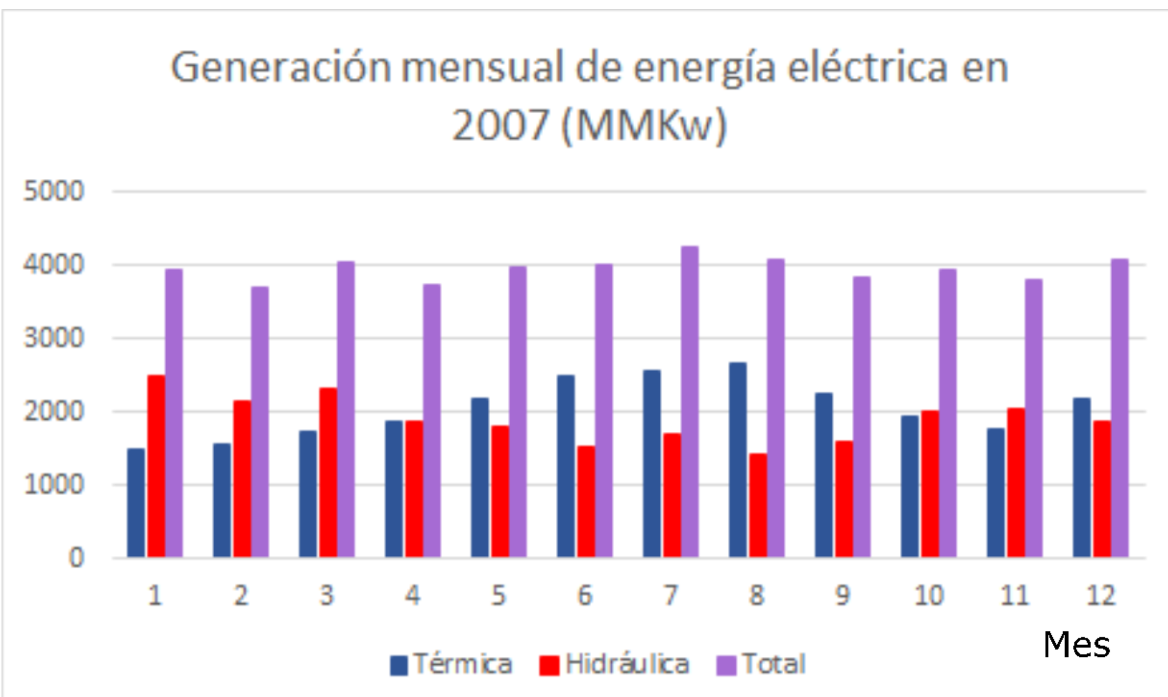


Ahora mostramos un gráfico de barras de ambas energías generadas, por mes.

Vemos que a medida que aumenta una, disminuye la otra.

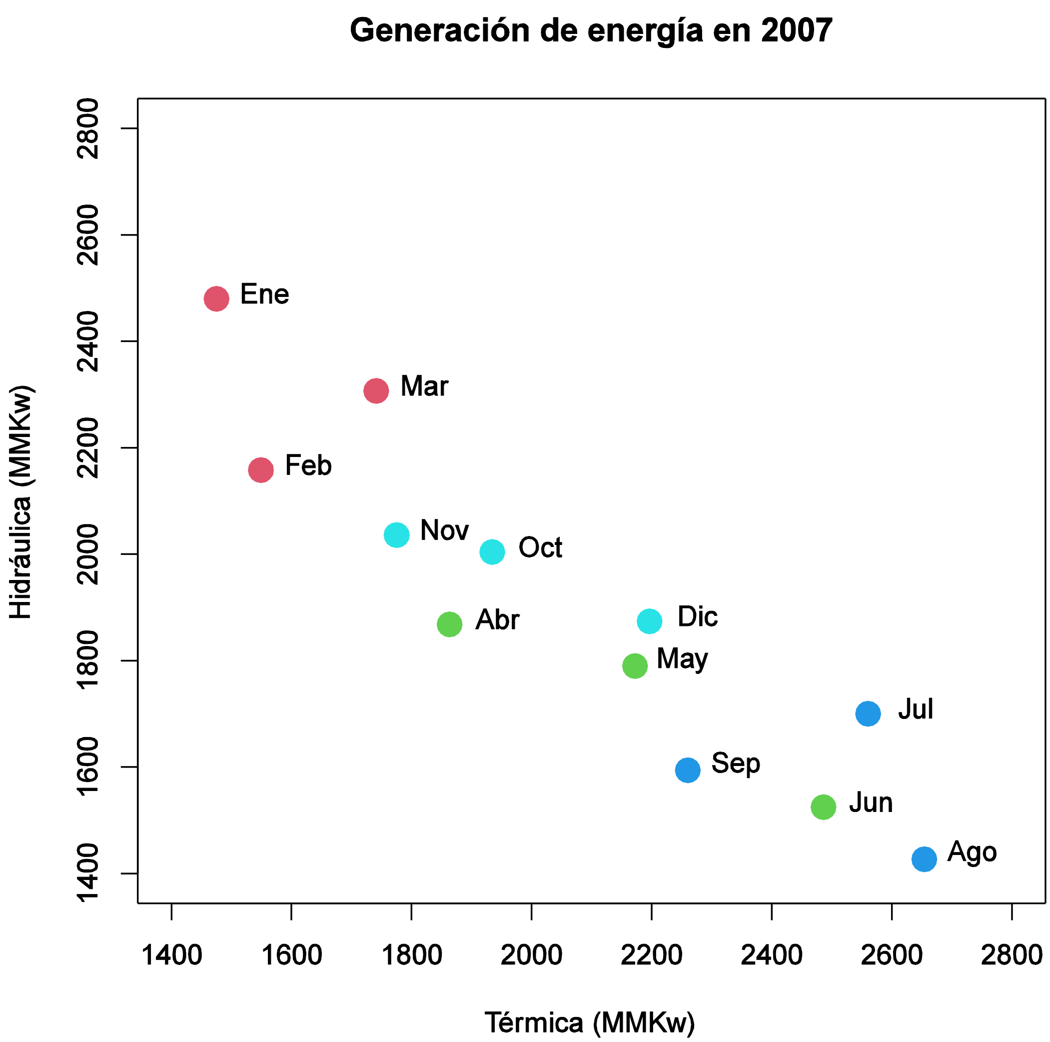


A continuación, el mismo gráfico, con los totales por mes, que se ven casi constantes a través de los meses.



Usaremos R para producir un diagrama de dispersión similar al anterior, en que el eje horizontal (de las abscisas) tiene la generación por medios térmicos y el eje vertical (de las ordenadas) tiene la generación por medios hidráulicos, en millones de kilowatts.

El gráfico nos muestra lo que ya habíamos visto, que a medida que hay más generación hidráulica, hay menos generación térmica, y viceversa: hay una asociación negativa.



Hemos agregado colores distintos a cada trimestre. Esto nos muestra que de enero a marzo hay más generación hidráulica.

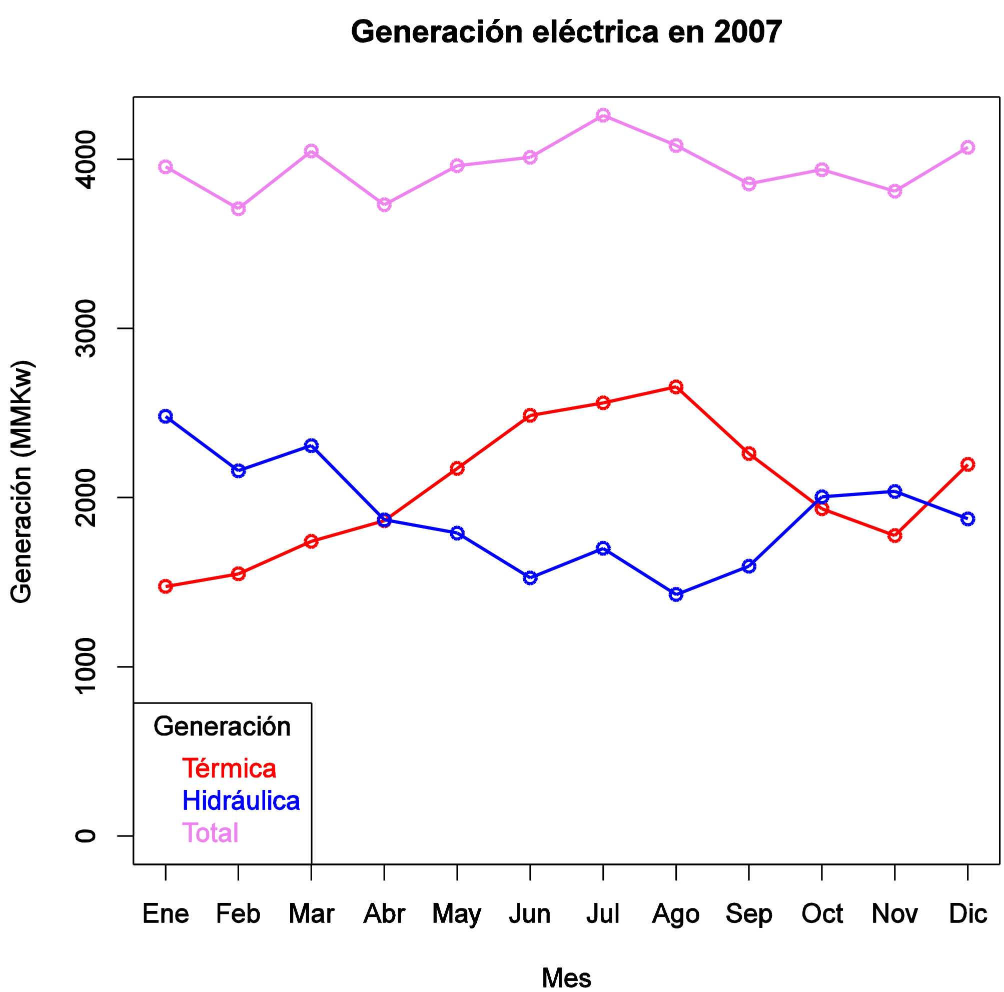
Puede ser porque el derretimiento de las nieves produce mayores caudales de agua en los ríos en que están las centrales eléctricas.

De junio a septiembre, fines de junio e invierno, hay más generación térmica, seguramente por las nevadas en las fuentes de los ríos.

Además, se aprecia que la relación es más o menos lineal. Y que la suma es aproximadamente constante.

El siguiente diagrama de dispersión tiene el mes en el eje horizontal y la generación de cada tipo y la suma en el eje vertical.

Por cada una se han unido los puntos para poder diferenciarlas con más facilidad.



En este gráfico se ve que mientras la generación hidráulica disminuye, la generación térmica aumenta, aunque la suma de ambas es aproximadamente constante, aumentando moderadamente en el mes de julio.

Observa que la escala vertical comienza desde cero. Si no partiera de cero, el gráfico daría una visión distorsionada.

Eso lo debe manejar el usuario, pues usualmente los programas fijan los límites de forma automática para que las curvas quepan justo dentro del área del gráfico.

EJEMPLO 26

Un grupo de 46 empresas de transporte terrestre, llevan registro de los gastos efectuados durante un año, en 8 rubros, en millones de pesos. Son los siguientes datos:

Las variables medidas a cada empresa son:

Id - Identificador de la empresa

RT - Revisión técnica

MU - Materiales y útiles

Tel - Telefonía

Pr - Primas de seguros

Mnt - Mantención

GPS - Servicios de GPS

GO - Gastos de Operación

IGP - Impuestos aduana, contribuciones, patentes

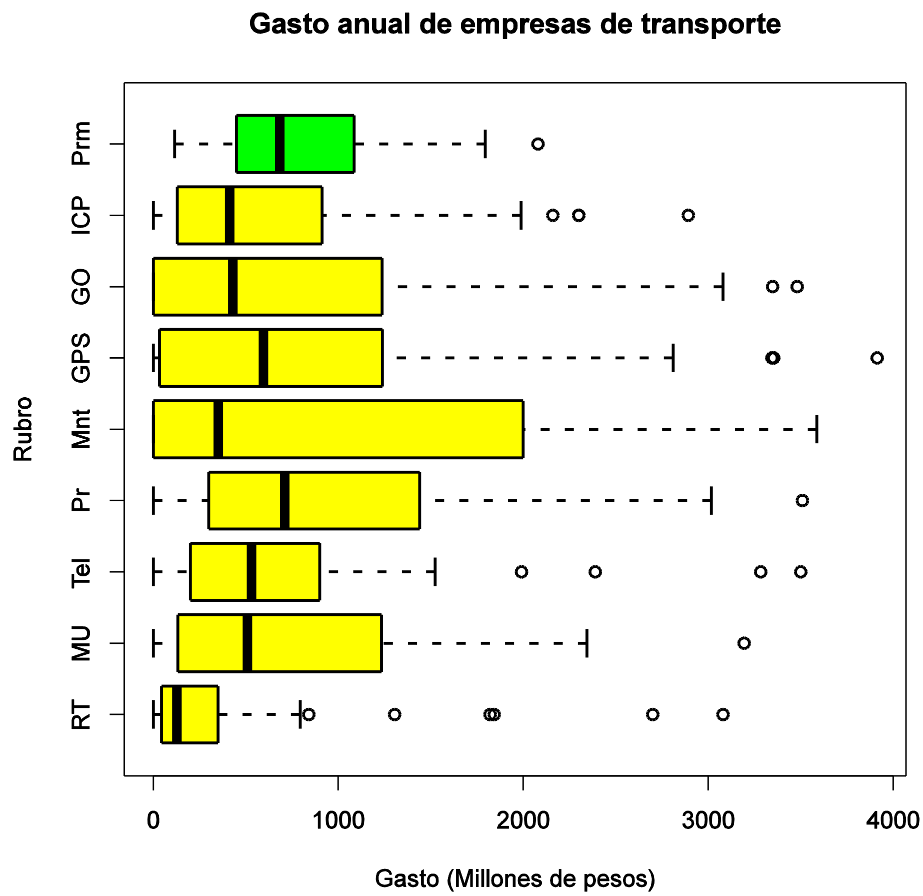
Prm – Promedio

Id	RT	MU	Tel	Pr	Mant	GPS	GO	ICP	Prm
1	166	2345	769	1115	0	0	2300	837	0
2	2700	900	0	2850	3489	500	654	1850	1618
3	82	1775	2389	534	737	2689	412	71	1086
5	1305	1380	1256	1654	2658	1239	0	2160	1457
6	0	7	405	0	0	1296	397	47	269
7	246	1763	294	2849	1346	3913	2005	1953	1796
8	201	1140	774	3509	0	0	702	322	831
9	41	630	397	195	0	628	3348	174	677
10	300	0	300	300	0	475	0	300	209
11	63	323	882	1040	76	0	458	138	373
12	1821	1123	894	1837	3200	2810	2060	2892	2080
13	100	200	180	200	0	34	420	147	44
14	0	963	1991	758	1215	1122	126	189	796
15	200	0	456	300	0	1266	445	0	333
16	94	237	900	0	2418	345	0	130	516
17	100	50	200	500	2000	300	0	100	406
18	298	265	432	778	711	720	345	456	501
19	24	1254	0	720	0	1728	0	267	499
20	93	186	756	347	782	0	2348	893	676
21	40	557	1500	1595	0	0	1234	205	641
22	100	50	200	500	2000	300	126	100	422
23	450	1234	300	200	0	541	800	210	467

Id	RT	MU	Tel	Pr	GPS	GO	ICP	Prm	Mant
24	45	236	0	175	0	0	456	17	116
25	520	30	0	200	0	234	2225	1089	537
26	3080	720	372	1000	0	1000	0	900	884
27	138	2013	580	2732	358	3354	2539	713	1553
28	1843	515	1089	886	1225	1782	2710	1892	1493
29	140	3194	1182	500	267	677	0	450	801
30	756	1129	1524	3017	0	0	2467	524	1177
31	0	2166	0	0	342	127	13	0	331
32	1	1377	141	619	75	1660	636	1083	699
33	80	133	982	504	3587	850	1237	0	922
34	720	240	0	347	834	560	0	1595	537
35	36	101	909	1905	3219	400	1912	646	1147
36	36	327	900	0	1860	1112	0	50	536
37	139	950	85	2594	3094	3344	1097	913	1527
38	23	11	234	486	135	8	0	46	118
39	75	500	675	875	2345	783	0	250	688
40	116	234	0	400	3218	794	0	152	614
41	290	2040	3500	1440	3467	1127	713	490	1633
42	52	0	630	1180	0	651	3481	402	800
43	841	1188	3284	700	0	0	3079	780	1234
44	350	300	481	2000	0	0	0	459	449
45	15	0	720	120	1800	76	341	96	396
46	166	2345	769	1115	0	0	0	2300	837

Construiremos diagramas de cajón con bigotes a partir de estos datos, de modo de poder comparar los gastos en los diferentes rubros.

Se muestra en verde el diagrama de cajón correspondiente a los promedios por empresa.



Podemos observar que, en general, en todos los rubros hay bastante **asimetría** en la forma como se distribuyen los gastos, y que la mayoría de los datos se concentran hacia valores bajos, con algunos pocos valores altos.

En todos los rubros, excepto en Mantención, hay **valores extremos**, todos grandes.

Si tratamos de pesquisar estos valores extremos en la tabla de datos, veremos que son gastos extraordinariamente grandes y que, en general, se trata de un caso distinto por cada empresa.

Sólo hay tres empresas con gastos extremos en dos rubros. Una de ellas es la identificada con el número 12, que es responsable del único valor extremo que aparece en el promedio.

EJEMPLO 27

Se tiene los siguientes datos que corresponden al número de accidentes laborales en el mes de enero de 2021, en cada región, por hombres, mujeres y totales.

Estos datos se extrajeron del sitio web del Instituto Nacional de Estadísticas (INE)

Se agregaron las poblaciones de las regiones, de Hombre, Mujeres y Totales, tomadas del Censo 2017.

Las variables registradas son:

Región - el nombre de cada región.

Num.Región - su número.

acc.Hom - número de Hombres afectados por accidentes de trabajo.

acc.Muj - número de Mujeres afectados por accidentes de trabajo.

acc.tot - número Total de afectados por accidentes de trabajo.

pob.Hom - población de Hombres.

pob.Muj - población de Mujeres.

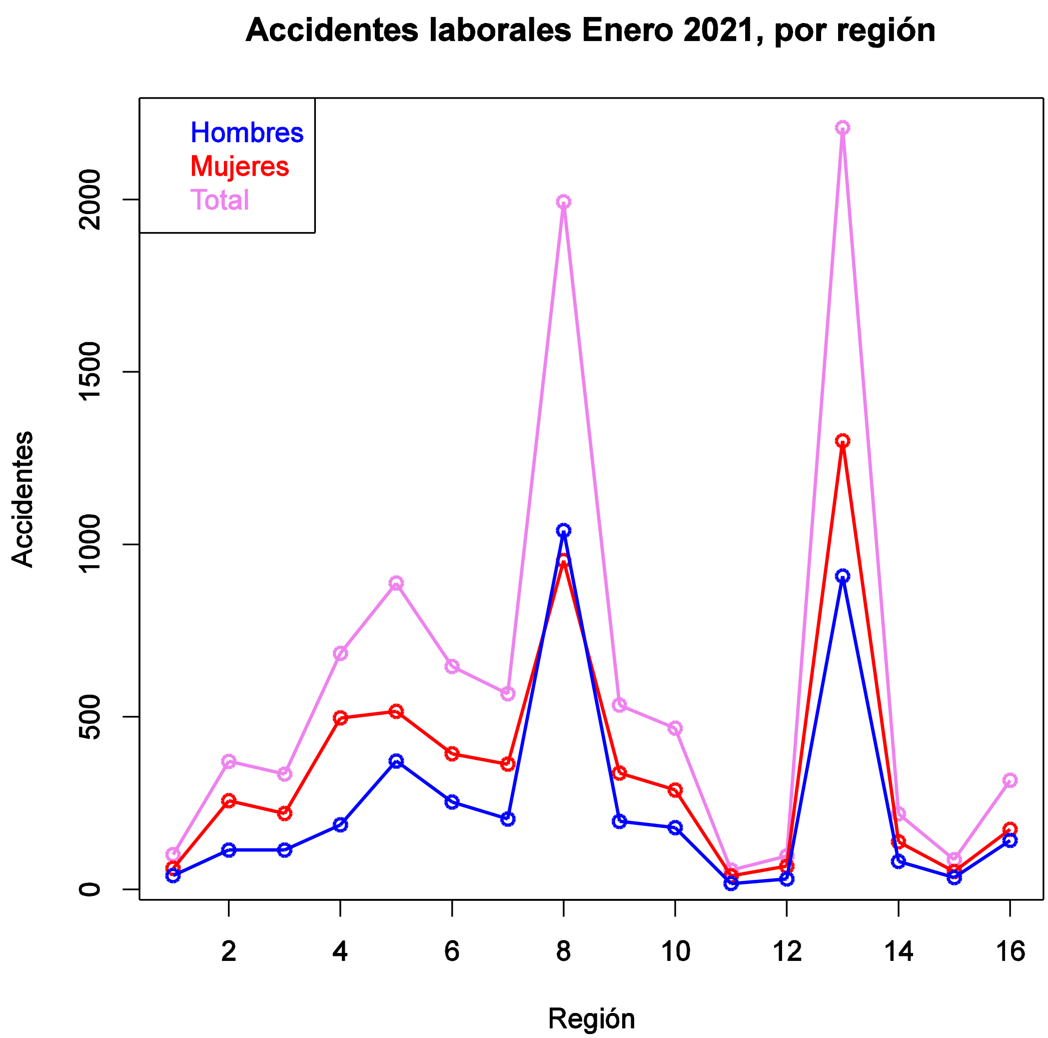
pob.tot - población Total.

Los datos son los siguientes:

Región	Num.región	acc.Hom	acc.Muj	acc.tot	pob.Hom	pob.Muj	pob.tot
Tarapacá	1	40	61	101	167793	162765	330558
Antofagasta	2	114	257	371	315014	292520	607534
Atacama	3	114	220	334	144420	141748	286168
Coquimbo	4	187	497	684	368774	388812	757586
Valparaíso	5	372	516	888	880215	935687	1815902
O'Higgins	6	253	393	646	453710	460845	914555
Maule	7	204	363	567	511624	533326	1044950
Biobío	8	1040	953	1993	750730	806075	1556805
La Araucanía	9	197	337	534	465131	492093	957224
Los Lagos	10	179	288	467	409400	419308	828708
Aysén	11	17	39	56	53647	49511	103158
Magallanes	12	30	67	97	85249	81284	166533
Metropolitana	13	908	1300	2208	3462267	3650541	7112808
Los Ríos	14	81	138	219	188847	195990	384837
Arica y Parinacota	15	34	52	86	112581	113487	226068
Ñuble	16	142	174	316	232587	248022	480609

Haremos un gráfico de los accidentes versus la región, están ordenadas por número, separando en tres variables: accidentes de Mujeres, de Hombres y Total.

Uniremos los puntos correspondientes a una misma variable por líneas con el objeto de visualizar mejor cada variable por separado.



Podemos observar que las tres líneas quebradas siguen el mismo patrón. Sin embargo, la de las Mujeres está más arriba en todos los puntos, salvo en la región 8 (Biobío), en que es ligeramente inferior.

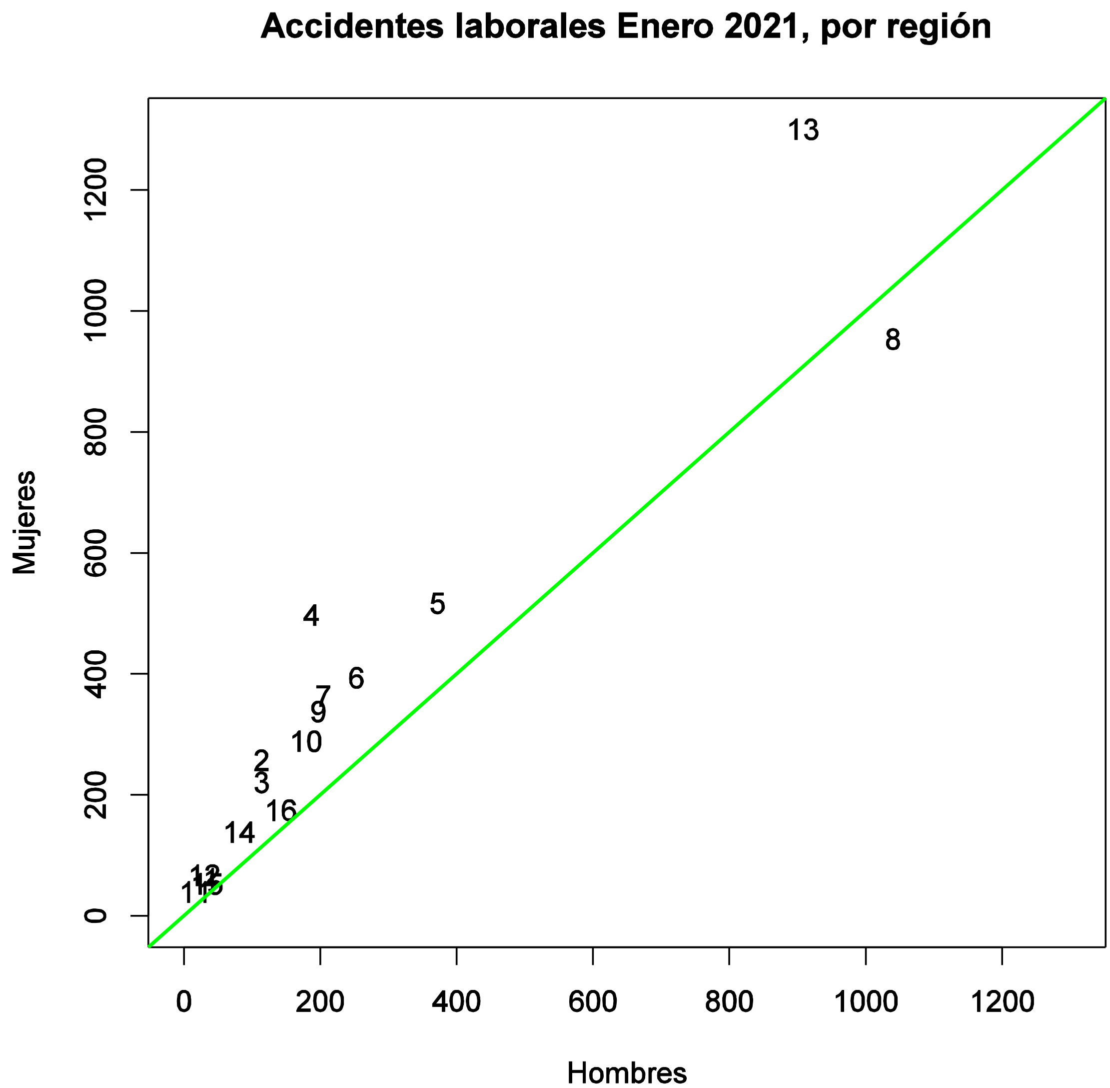
También vemos que las Regiones 8 (Biobío) y 13 (Metropolitana), tienen notoriamente más accidentes que el resto de las regiones.

En el caso de la región 13 es claro que se debe a que es una Región notablemente más poblada que las demás.

Sin embargo, la Región 5 (Valparaíso) tiene una mayor población que la 8, pero tiene menos accidentes. Aunque sí tiene más que las restantes Regiones.

A continuación, haremos un gráfico de dispersión del total de accidentados Mujeres versus el total de accidentados Hombres, identificando cada punto con el número de la Región.

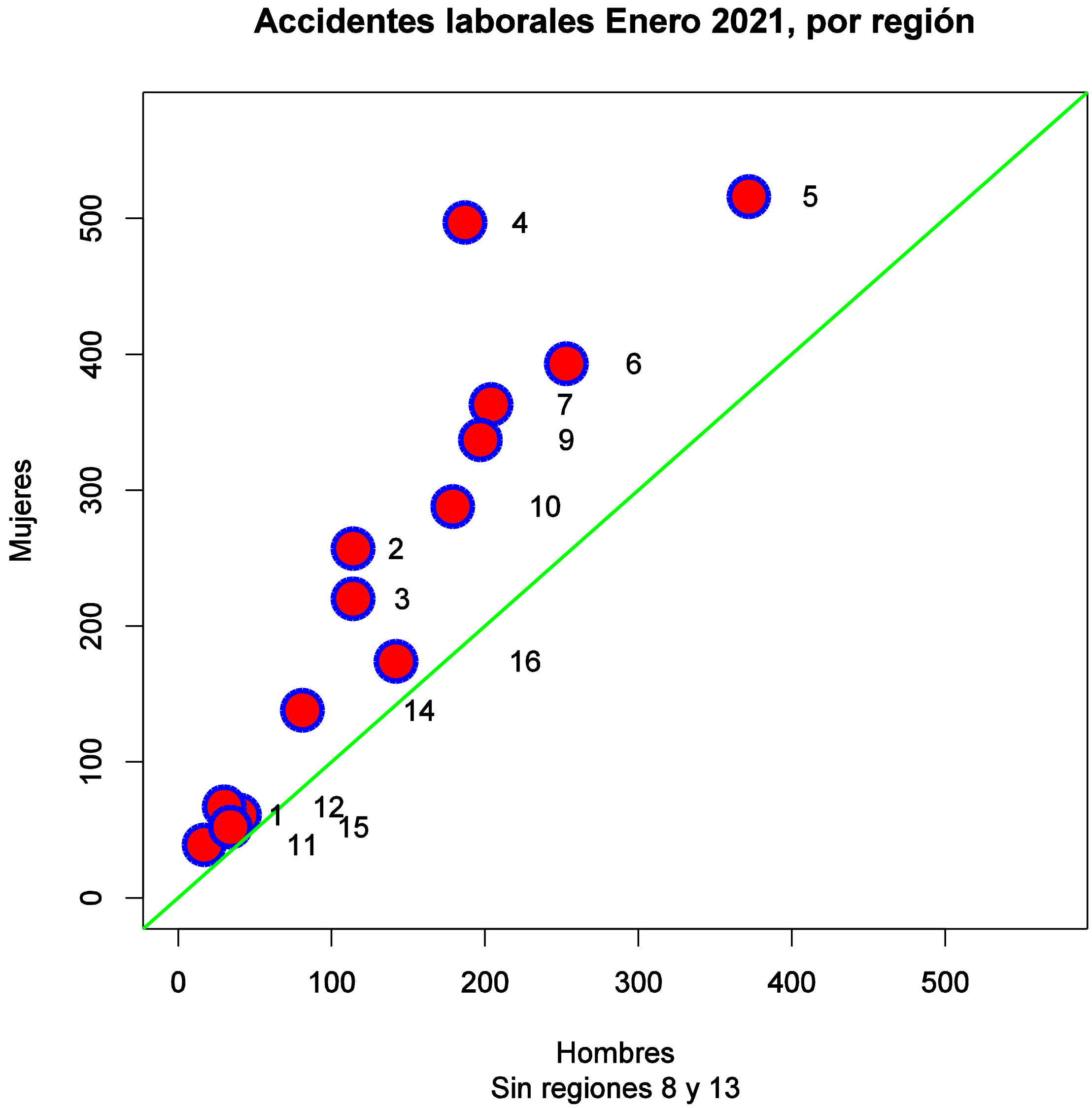
Se trazó una línea diagonal verde, que representaría la igualdad entre accidentes de Hombres y Mujeres.



Aquí se observa claramente que los puntos aparecen sobre la recta diagonal, lo que ratifica el hecho que el número de accidentes de Mujeres es superior al de los Hombres, salvo en el caso de la Región 8.

Además, el gráfico destaca el hecho que las Regiones 8 y 13 tienen números mucho más grandes de accidentes.

Con el objeto de observar con más claridad el comportamiento de las demás regiones, haremos el mismo gráfico de dispersión, pero sin incluir las Regiones 8 y 13, de modo de agrandar la escala.



En este gráfico vemos que, entre estas, la Región 5 es la de más accidentados, y la 4 es la que tiene notoriamente más accidentadas Mujeres que Hombres.

Las con menos accidentados son las Regiones 1 (Tarapacá), 11 (Aysén), la de menor población, 12 (Magallanes) y 15 (Arica y Parinacota).

Interesa visualizar los accidentados estandarizados, es decir, como si todas las Regiones tuvieran la misma población.

Para eso definimos tres Tasas:

Accidentados Hombres por cada 100.000 habitantes Hombres.

Accidentadas Mujeres por cada 100.000 Mujeres.

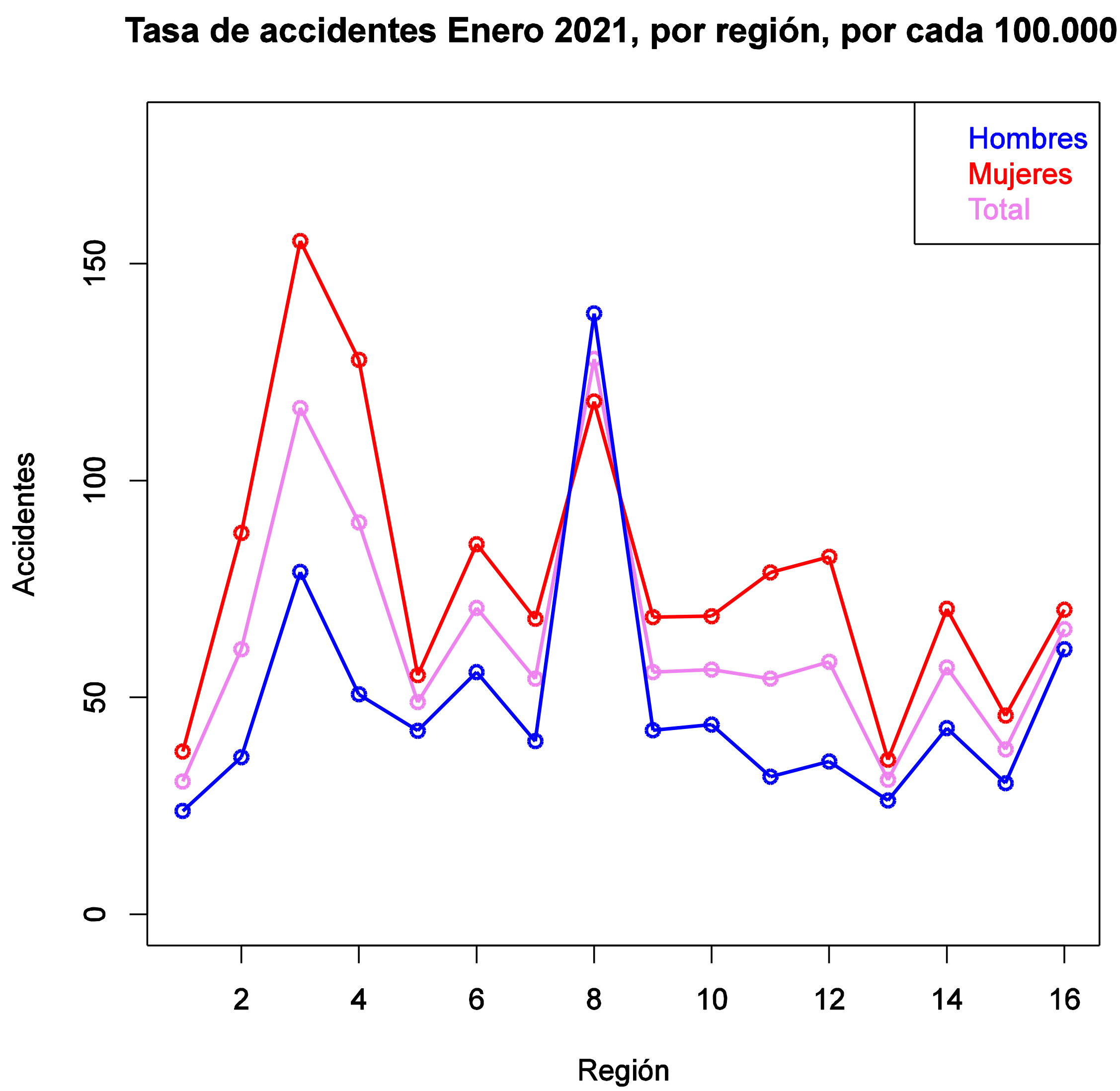
Accidentados totales por cada 100.000 habitantes (Hombres y Mujeres).

El cálculo de cada tasa se hace dividiendo el valor de accidentados por la población respectiva y multiplicando por 100.000. Lo redondeamos a un decimal.

Los valores obtenidos son los siguientes:

Num.región	tasa.Hom	tasa.Muj	tasa.tot
1	23.8	37.5	30.6
2	36.2	87.9	61.1
3	78.9	155.2	116.7
4	50.7	127.8	90.3
5	42.3	55.1	48.9
6	55.8	85.3	70.6
7	39.9	68.1	54.3
8	138.5	118.2	128.0
9	42.4	68.5	55.8
10	43.7	68.7	56.4
11	31.7	78.8	54.3
12	35.2	82.4	58.2
13	26.2	35.6	31.0
14	42.9	70.4	56.9
15	30.2	45.8	38.0
16	61.1	70.2	65.7

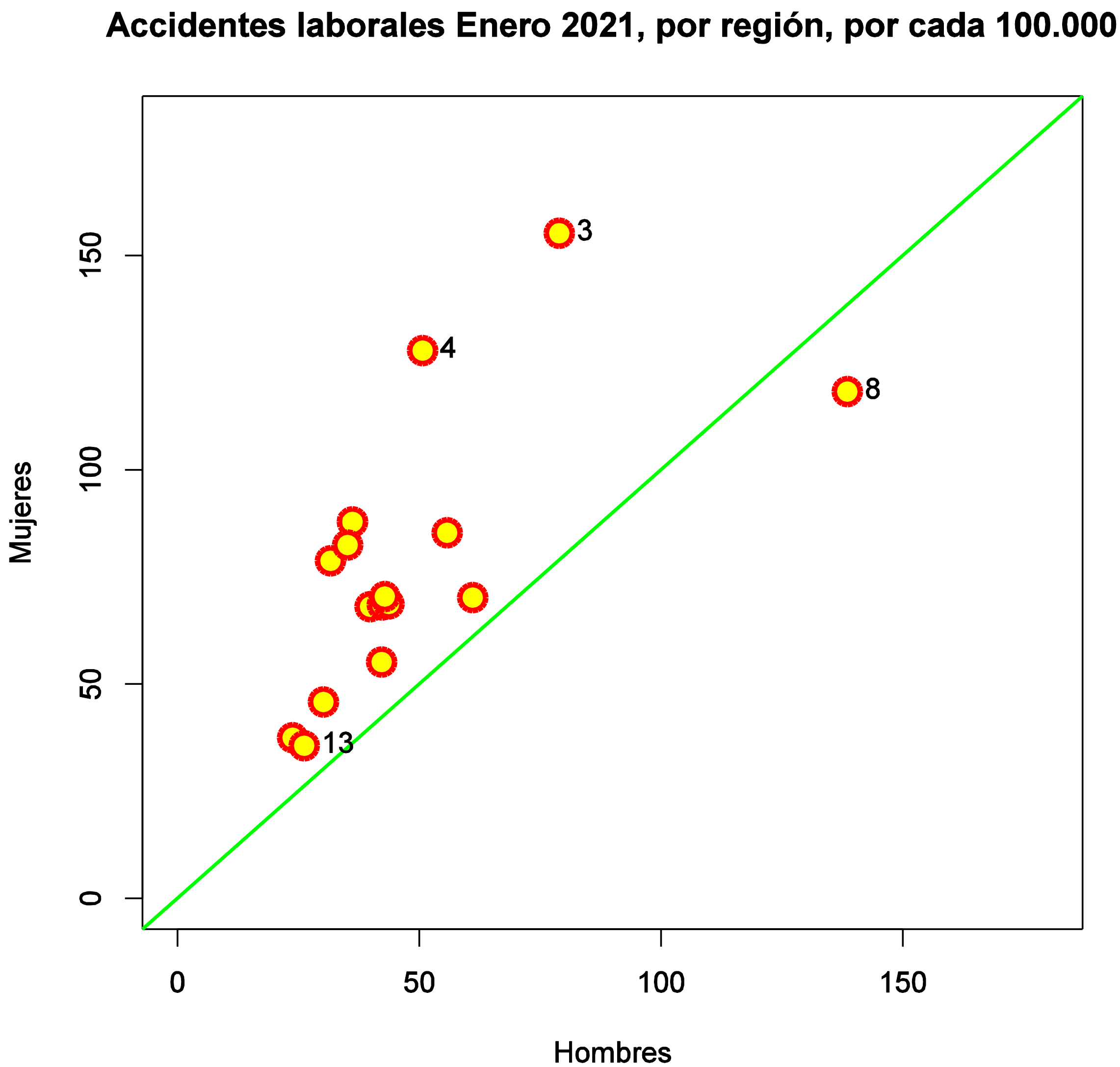
Con estos datos construimos un gráfico de Tasa de accidentes versus la región, separados en Mujeres, Hombres y Total. Los puntos de cada variable unidos por rectas.



Podemos ver que ahora las tasas de accidentes son mayores para las regiones 3 (Atacama), 4 (Coquimbo) y 8 (Biobío).

Pero también observamos que, a diferencia de los Totales, las Tasas son mayores para las Mujeres que para los Hombres y que para los Totales, en todas las Regiones excepto la 8.

Finalmente, con estos datos construimos un gráfico de dispersión de la tasa de accidentadas Mujeres versus la tasa de accidentados Hombres. Se incluye las regiones 8 y 13 que se habían eliminado.



En este gráfico se puede ver que cambia la posición relativa de los puntos.

En efecto, podemos ver que la Región 13 (Metropolitana), que era la de mayor cantidad de accidentes, ahora pasó a ser la de menor Tasa, accidentados por 100.000 habitantes, similar a la región 1 (Tarapacá).

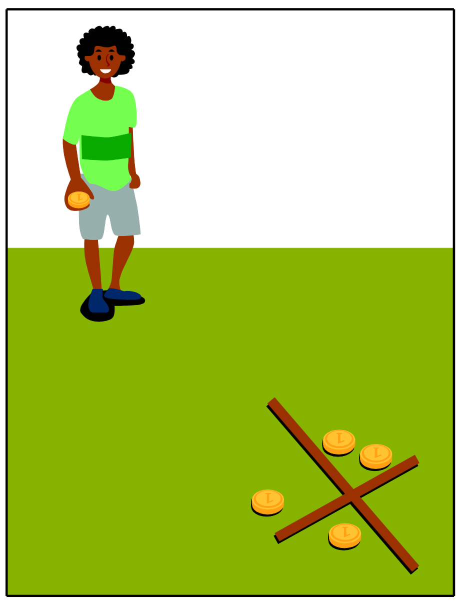
Mientras que la Región 8 se mantiene teniendo un valor más grande, en los accidentados Hombres.

Junto a las Regiones 3 (Atacama) y 4 (Coquimbo), que tienen tasas altas en el caso de Mujeres.

ACTIVIDAD PRÁCTICA

Traza dos líneas rectas perpendiculares en el suelo. Párate a unos tres a cuatro metros detrás, en la dirección de una de las líneas perpendiculares.

Lanza monedas tratando de apuntarle el punto en que se cruzan las dos líneas.



Mide y registra la distancia, en centímetros, desde donde cayó la moneda hasta cada una de las dos líneas.

Esto lo repites varias veces.

Pueden participar todos los alumnos del curso, divididos en **equipos**, de modo de juntar una cantidad grande de pares de distancias.

Con estos datos puedes construir un gráfico de dispersión, que te permitirá determinar en qué dimensión hay más error, la dirección del lanzamiento o la dirección perpendicular.

En el gráfico, separa los puntos según el equipo participante, usando diferentes colores.

Obtén algunas conclusiones.

EJEMPLO 28

Los datos siguientes corresponden a producción de 21 productos agrícolas en los períodos 2019-2020 y 2020-2021, en quintales (1 quintal=100 kilos).

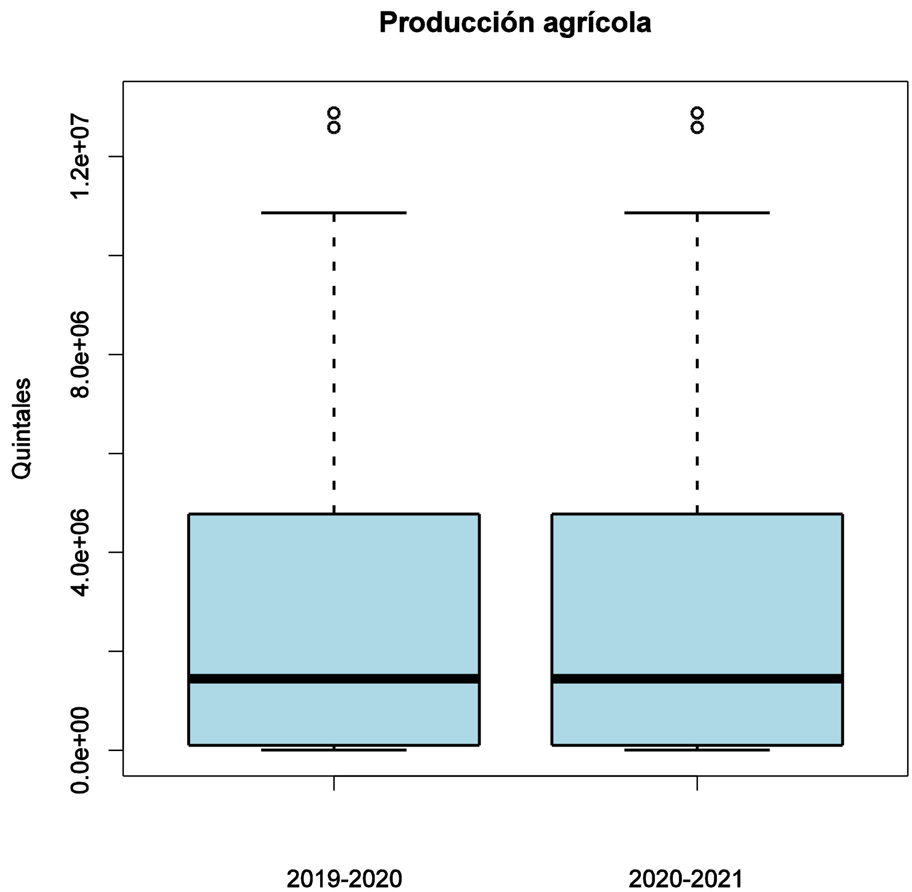
En la tabla aparece el cultivo, la categoría a la que pertenece y la subcategoría.

Nuestro objetivo es comparar la producción en los dos períodos.

La fuente de los datos es el Instituto Nacional de estadísticas, INE.

num	Cultivo	prod19-20	prod20-21	categoría	Sub categoría	ncat
1	Trigo Harinero	10861401	12033828	cereales	trigo	1
2	Trigo Candeal	1448483	1502251	cereales	trigo	1
3	Cebada Cervecera	1586470	1164548	cereales	cebada	1
4	Cebada Forrajera	345189	410764	cereales	cebada	1
5	Avena	4773956	5252446	cereales	avena	1
6	Maíz Consumo	5658838	7719603	cereales	maíz	1
7	Maíz Semilla	271043	218615	cereales	maíz	1
8	Arroz	1696965	1460851	cereales	arroz	1
9	Triticale	928634	551828	cereales	tricale	1
10	Poroto	91774	174910	leguminosas	poroto	2
11	Lenteja	5850	5717	leguminosas	lenteja	2
12	Garbanzo	1853	3116	leguminosas	garbanzo	2
13	Papa	12881536	9945078	papas	papas	3
14	Raps	1535334	1405858	industriales	raps	4
15	Maravilla	34097	21433	industriales	maravilla	4
16	Lupino Amargo	98776	160964	industriales	lupino	4
17	Otros Lupinos	200870	209531	industriales	lupino	4
18	Remolacha azucarera	12590481	7462724	industriales	remolacha	4
19	Tabaco	64389.	39248	industriales	tabaco	4
20	Tomate Industrial	5234810	6768235	industriales	tomate	4
21	Achicoria Industrial	1939388	2674475	industriales	achicoria	4

Primero que nada, haremos dos diagramas de cajón con bigotes de los dos períodos.



Podemos observar que, en general, se ven parecidas las producciones en ambos períodos.

También se observa una fuerte asimetría superior, con dos valores extremos.

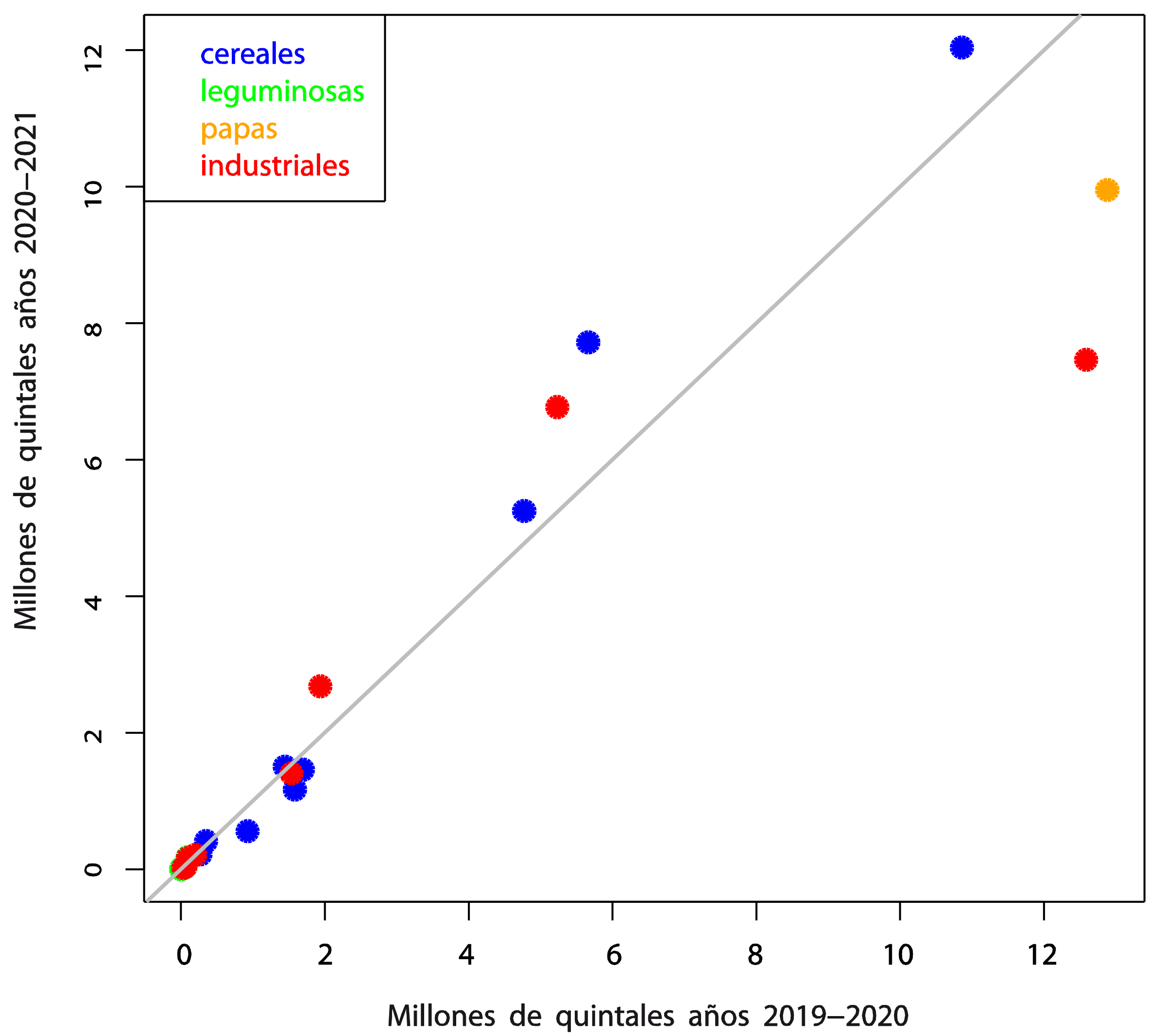
Estos son la Papa y la Remolacha Azucarera, ambas exceden los 12 millones de quintales.

Ahora haremos un diagrama de dispersión. Pondremos la producción en 2019-2020 en el eje horizontal y la producción 2020-2021 en el eje vertical.

Esto nos permitirá comparar los dos períodos.

Hay cuatro categorías: Cereales, Leguminosas, Papas e Industriales. Los puntos correspondientes a cada una los representaremos en colores diferentes.

Producción agrícola



Hemos trazado una línea diagonal que representa los puntos en que ambas variables toman los mismos valores.

Si un punto está sobre la diagonal significa que hubo un aumento en la producción 2020-2021 respecto del período anterior.

Si el punto está debajo de la diagonal, hubo una disminución.

Hay que aclarar que todo lo que se ve en un gráfico, también se puede deducir directamente de los datos.

¿Cuál es la ventaja del gráfico? Que se ve todo con mucho mayor claridad que de una tabla de datos. El gráfico da información menos precisa, pero más fácil de entender.

El problema que tenemos es que nos encontramos con que hay algunos puntos dispersos, pero hay una gran concentración de puntos en torno a los valores bajos, que no permiten ver el detalle.

Entonces haremos lo siguiente: Dividiremos los datos en tres grupos, de acuerdo a la producción en el período 2019-2020:

El grupo de Baja Producción, con los 7 productos que presentan producción más baja en el período 2019-2020.

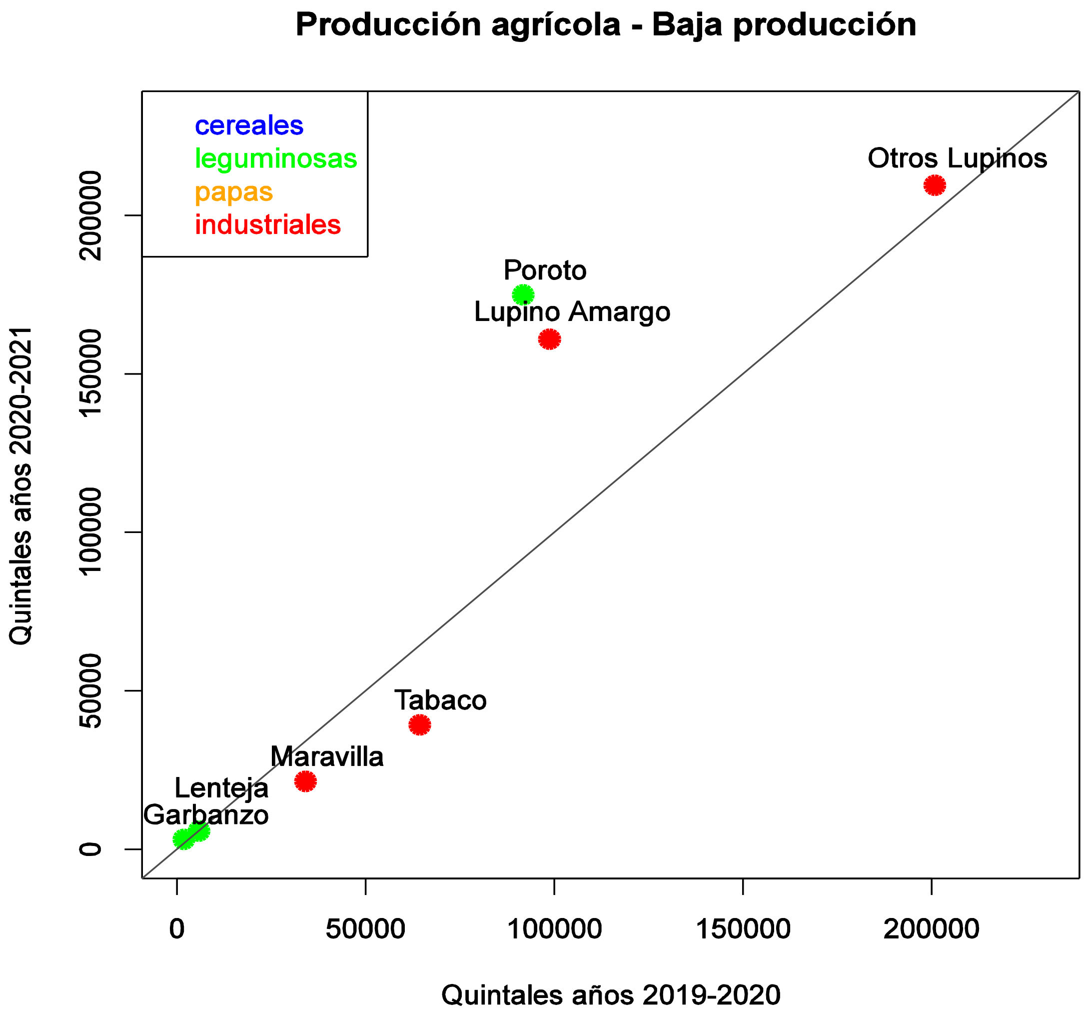
El grupo de Producción Media, los siguientes 7 productos.

El grupo de Alta Producción, los 7 con la producción más alta en 2019-2020.

Y representaremos cada grupo en un gráfico separado.

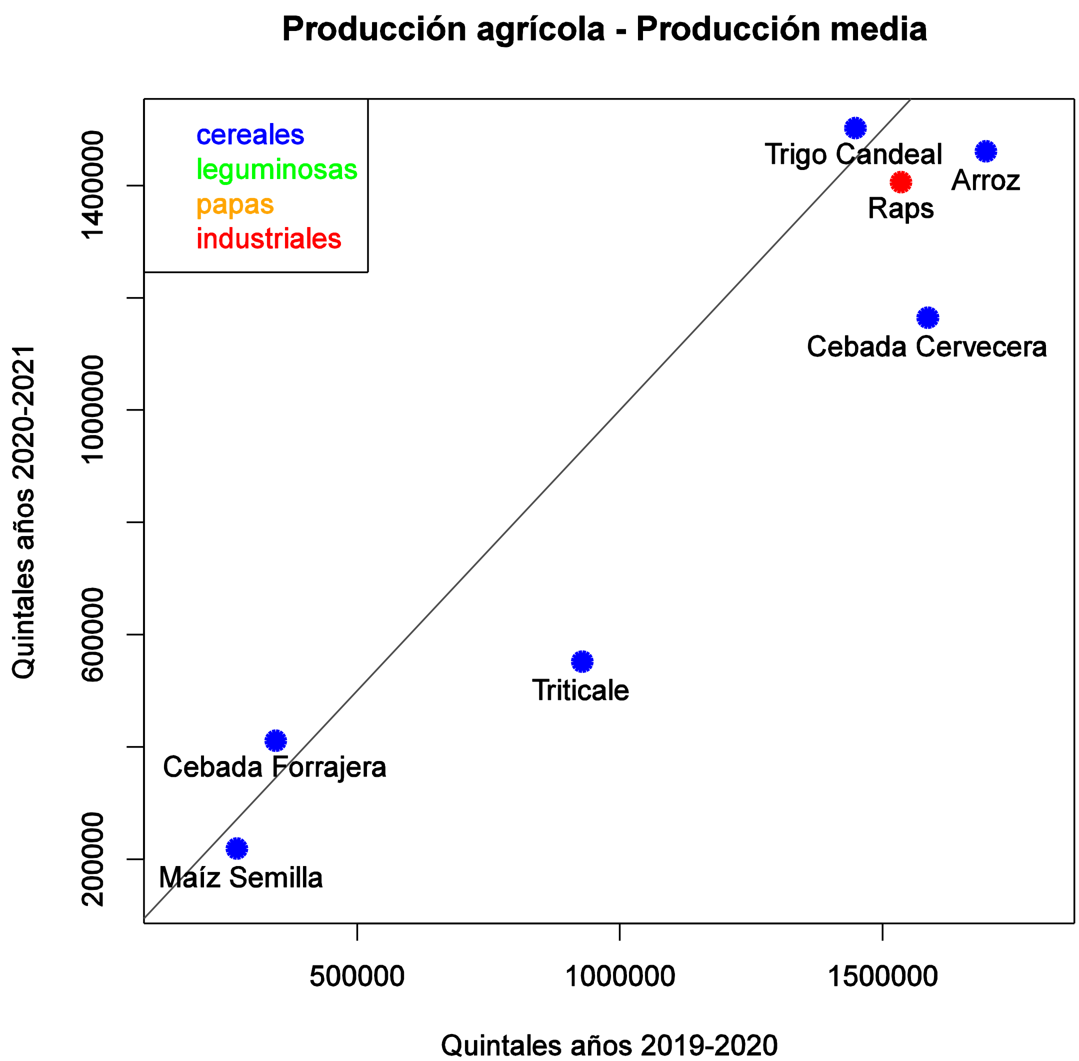
Esto es como hacerle un zoom al gráfico anterior.

Los resultados son los siguientes, partiendo por Baja Producción, e individualizando cada producto:



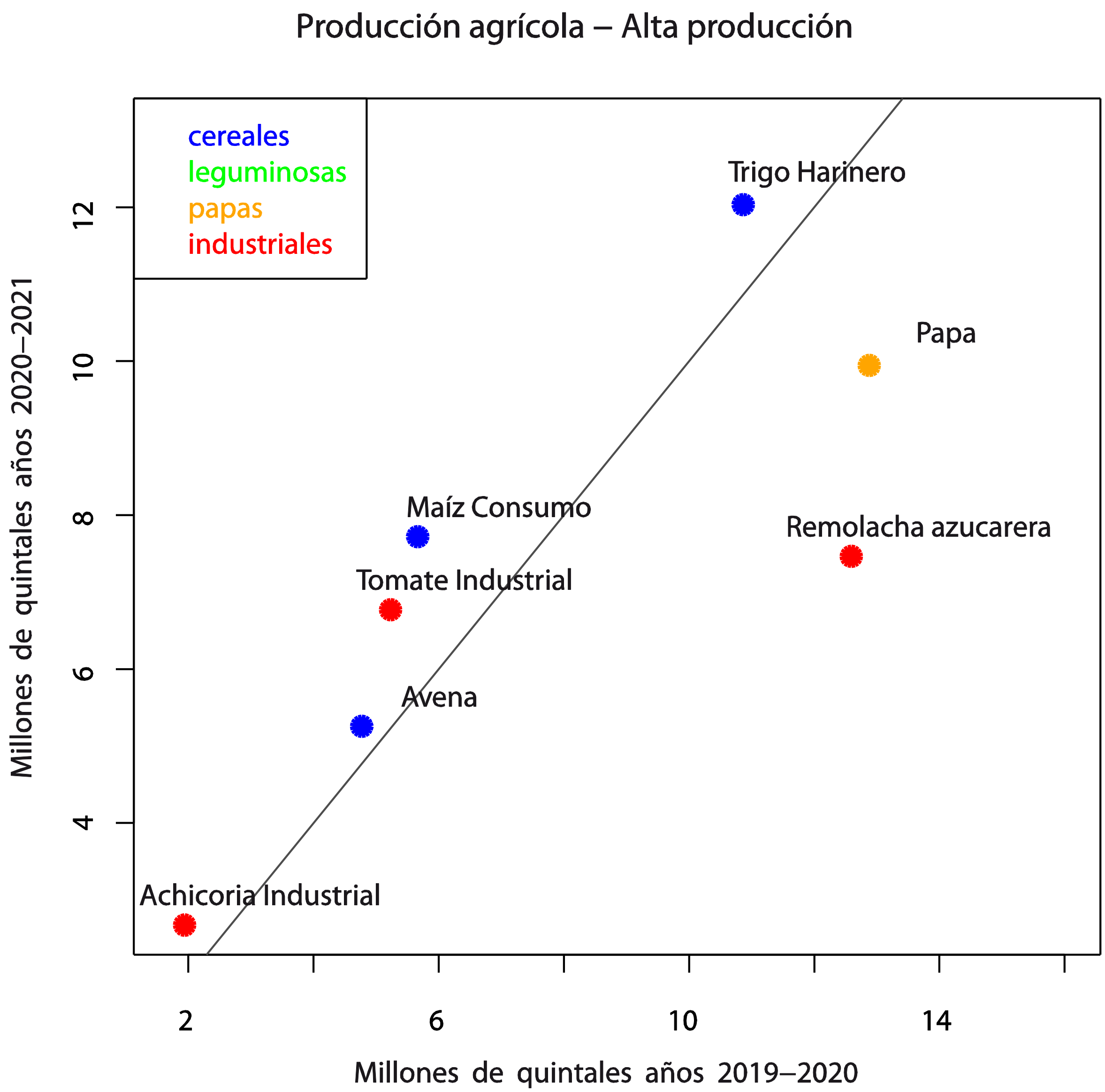
Se ven las tres Leguminosas, en verde, y algunos productos industriales, en rojo.

Están cerca de lo línea diagonal, excepto Poroto y Lupino Amargo, que muestran un aumento en su producción.



En la Producción Media aparece un producto industrial, el Raps. Todos los demás son Cereales, representados en azul.

Algunos de ellos, Cebada Cervecera, Arroz y Triticale con una disminución en su producción en 2020-2021 respecto del período anterior. Raps también, pero con una disminución menor.



En este gráfico, de Alta Producción, aparece la Papa y la Remolacha Azucarera, con disminuciones en su producción. Algunos de los demás productos tienen aumentos menores en la producción.

Si se desea hacer comparaciones entre los tres gráficos, se debe tener en cuenta que tienen escalas diferentes.

Las marcas del eje vertical del gráfico de baja producción van de 0 a 200.000 quintales.

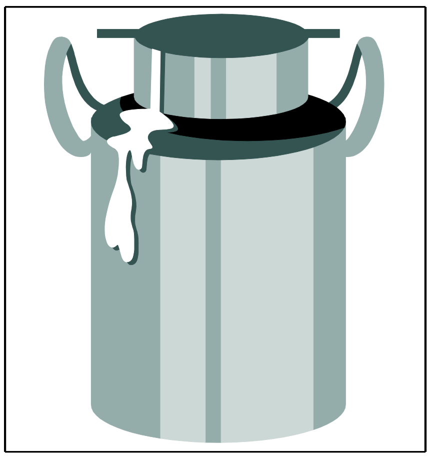
Las de Producción media, de 200.000 a 1.400.000 quintales.

Las de Alta producción van de 4 millones a 12 millones de quintales.

Por lo tanto, una distancia de un centímetro en uno de los tres gráficos no corresponde a lo mismo que un centímetro en otro.

EJERCICIOS

19) La tabla de datos siguiente muestra la producción de leche en la Región de Los Lagos, en el cuarto trimestre de los años 2018, 2019 y 2020.



Las variables son:

A - año

M - mes

NProc - número de productores

VolTot - volumen total producido (miles de litros Mlt)

Propia - producción propia (Mlt)

Adquirida - producción adquirida (Mlt)

NPCompran - cantidad de plantas que compran leche

NProv - número de proveedores de leche.

Año	Mes	NProc	VolTot	Propia	Adquirida	NPCompran	NProv
2018	Enero	18	4830	150	4680	11	141
2018	Febrero	18	3706	135	3570	11	139
2018	Marzo	18	3743	135	3608	11	133
2018	Abril	16	3414	109	3305	10	130
2018	Mayo	16	3210	103	3107	10	132
2018	Junio	16	2891	100	2791	10	125
2018	Julio	17	2934	81	2853	11	127
2018	Agosto	17	3582	91	3491	12	130
2018	Septiembre	17	3904	87	3817	12	129
2018	Octubre	19	4856	134	4722	11	134
2018	Noviembre	19	5356	124	5232	11	135
2018	Diciembre	19	5130	113	5017	12	131
2019	Enero	17	4751	152	4598	11	134
2019	Febrero	17	4084	141	3943	11	139
2019	Marzo	17	3710	125	3585	12	144
2019	Abril	16	2926	116	2810	11	129
2019	Mayo	16	3025	106	2919	11	127
2019	Junio	16	2773	109	2664	11	120

Año	Mes	NProc	VolTot	Propia	Adquirida	NPCompran	NProv
2019	Julio	17	2603	80	2523	11	112
2019	Agosto	17	2846	63	2782	11	105
2019	Septiembre	17	3536	64	3472	12	105
2019	Octubre	16	4359	125	4234	11	108
2019	Noviembre	16	4747	124	4623	11	114
2019	Diciembre	16	4599	130	4469	11	111
2020	Enero	16	4220	124	4096	11	111
2020	Febrero	16	3764	115	3649	11	111
2020	Marzo	16	3205	107	3098	11	109
2020	Abril	15	2518	71	2448	10	109
2020	Mayo	15	2772	69	2703	10	109
2020	Junio	15	2389	61	2329	10	107
2020	Julio	15	2317	54	2263	9	106
2020	Agosto	15	2720	55	2665	10	113
2020	Septiembre	15	3491	68	3423	10	118
2020	Octubre	15	4423	101	4322	10	114
2020	Noviembre	15	4694	103	4591	10	113
2020	Diciembre	15	4392	70	4322	10	118

a) Haz un análisis similar a lo del Ejemplo 28, de la producción agrícola, relacionando las variables **Número de productores** y **Volumen total producido**. De cada gráfico obtén conclusiones.

b) Lo mismo que en a), pero con las variables **Producción propia** y **Producción adquirida**.

c) Lo mismo que en a), pero con las variables **Cantidad de plantas que compran leche** y **Número de proveedores de leche**.

APÉNDICE 1 - Uso de Microsoft Excel

Las planillas Excel de los ejemplos las puedes encontrar en el sitio web del autor

www.jorgegalbiati.cl

o bien puedes escanear este código QR.

Para usar las funciones estadísticas de Excel se ocupa insertar función, que se muestra con el símbolo f_x.

Presionando f_x aparecerá un submenú con la lista de funciones disponibles.

Para gráficos se va a

Insertar

En el submenú **Gráficos** se elige el gráfico a usar.

Para procedimientos estadísticos más complejos se debe recurrir a

Datos

Análisis de Datos. Al presionarlo aparecerá un submenú con la lista de procedimientos disponibles.

Podría ser que **Análisis de Datos** no esté habilitado. En ese caso se debe hacer lo siguiente, una sola vez:

Archivo

Opciones

Complementos

Complementos de Excel

Análisis de datos

Con eso queda habilitada la opción **Análisis de Datos** en **Datos** todas las veces que ingresemos a Excel.

EJEMPLO 12

Seleccionar las dos columnas de datos, sin incluir los totales, pero sí incluyendo los encabezados superiores y los de la izquierda.

Luego posicionarse en INSERTAR, y de ahí seleccionar el tipo de gráfico que se quiere.

Aparecen dibujados diferentes tipos de gráfico.

En este caso elegir INSERTAR GRAFICO COLUMNAS O DE BARRAS.

De entre los que aparecen, seleccionar COLUMNA EN 2D. Automáticamente construirá el gráfico requerido, a excepción del título.

Posicionarse en la frase TÍTULO DEL GRÁFICO, presionar dos veces el botón derecho.

Eso permitirá escribir el título que uno quiere.

EJEMPLO 13

Como en el Ejemplo 12, pero seleccionar sólo la fila de frecuencias de la tabla de frecuencias.

Lo demás es igual que en el Ejemplo 12.

EJEMPLO 25

Los tres primeros gráficos de este Ejemplo fueron hechos con Excel.

Los otros se hicieron con R (ver Apéndice 2).

Para hacer un gráfico se deben seleccionar las columnas donde están los datos que se quiere representar, incluyendo los encabezados para que aparezcan identificados en el gráfico.

Luego posicionarse en INSERTAR, y de ahí seleccionar el tipo de gráfico que se quiere.

Para el primer gráfico se seleccionó un gráfico de dispersión, del tipo con puntos aislados.

Para los dos gráficos siguientes se seleccionó un gráfico de columnas verticales bidimensionales (2d).

Debe evitarse el uso de gráficos tridimensionales (3d) porque dan una idea distorsionada de la información contenida en los datos.

En la parte superior de los gráficos aparecerá un cuadro de texto con la frase "Título del gráfico".

Poniendo el cursor sobre ella y presionando el botón izquierdo dos veces se podrá escribir el título que se desee.

APÉNDICE 2 - Uso de R

Los programas R de los ejemplos los puedes encontrar en el sitio web del autor

www.jorgegalbiati.cl

o bien puedes escanear este código QR.

R es dos cosas a la vez: un lenguaje de programación y un paquete estadístico utilizado ampliamente por los estadísticos del mundo.

Nosotros lo usaremos fundamentalmente como paquete estadístico.

Es de código abierto y licencia libre.

Para instalarlo se debe ingresar al sitio www.r-project.org

Presionar **Download R**

Se pedirá elegir una de una lista de URL. Puede ser cualquiera, pero se recomienda una de Chile.

De ahí habrá que elegir una de tres opciones:

Download R for Linux (Debian, Fedora/Redhat, Ubuntu)

Download R for macOS

Download R for Windows

Luego presionar **Install R for the first time** y finalmente **Download R** (versión) for (**Linux**, **Mac** o **Windows**)

En breve tiempo estará disponible el programa de instalación, que se deberá correr y seguir las instrucciones. Todo el proceso no tarda más de 5 minutos.

Al correr R aparecerá la consola de R **(R Console**). En esta consola aparecerán los resultados. También se pueden ingresar comandos, que se ejecutarán con sólo presionar Enter. Entregará el resultado, pero el comando no será posible correrlo de nuevo, sin ingresarlo nuevamente.

Para poder conservar los comandos ingresados, se ingresa a

Archivo

Nuevo Script

Aparecerá un cuadro **Sin nombre - Editor R**. En él se podrán escribir comandos, editarlos, y guardarlo como archivo R para recuperarlo nuevamente en otra sesión de R. Se denomina un **Script**. Para recuperarlo:

Archivo

Abrir Script

Para ejecutar el script o parte de él:

Seleccionar lo que se quiere ejecutar e ir a:

Editar

Correr línea o seleccionar

Hagamos este sencillo ejercicio (lo que aparece después de # es un comentario no ejecutable):

Abrir un script nuevo. Ingresar:

```
M<-30   # define la constante M
v<-c(12,34,78,4,15)   # define el vector v
mean(v)   # promedio de los elementos de v
M *v   # multiplica los elementos de v por M.
b<-M-v   # sustrae los elementos de vector v a la constante M
# y los guarda en b para cálculos posteriores
b   # muestra el valor de b
```

Para ejecutarlo, se puede hacer por partes (respetando el orden de los comandos) o todo de una vez.

Tener en cuenta:

El programa distingue entre minúsculas y mayúsculas. No usar acentos ni la letra ñ.

Las constantes que se ingresan o que se calculan se mantienen en memoria mientras no se cierre el programa R.

Para ayuda en un comando, correr **help(comando)** y se desplegará la página correspondiente del manual.

A continuación, detallaremos el uso de R en algunos de los Ejemplos, para que te inicies en el conocimiento de este sistema computacional:

EJEMPLO 11

```
# Los comentarios van precedidos del símbolo #

# Ingresar los datos como un vector llamado alturas:

alturas<-c(146, 175, 147, 143, 170, 172, 177, 162, 160, 164, 185, 151,180,
161, 152, 171, 182, 163, 181, 169, 148, 184, 166, 157, 176, 156, 149, 178,
153, 168)

muestra1<sample(alturas,replace=TRUE)   # obtener muestra con reposicion

muestra1   # ver valores

mean(muestra1)   # calcular promedio

muestra2<-sample(alturas,replace=FALSE)   # obtener muestra sin reposicion

muestra2

mean(muestra2)

notas<-c (5.2, 4.6,....)  # ingresar las notas. Se usa punto decimal en lugar de
coma.

boxplot(notas,   # diagrama de cajón con los siguientes parámetros:

horizontal=TRUE,   # gráfico en forma horizontal

col="green",   # color de relleno

main="Notas quad Ciencias quad Sociales",   # título principal

xlab="Notas",   # título eje horizontal

lwd=2,   # ancho de líneas

ylim=c(1,7))   # límites eje horizontal
```

EJEMPLO 20

```
sueldos<-c (320,326,...)
length(x)
q<-c (0.25,0.75)  # cuartiles a calcular (en forma de fracción)
quantile(x,q)   # calcular cuartiles
median(x)   # calcular mediana
mean(x)   # calcular promedio
summary(x)   # este comando entrega todo lo anterior
tit<-'Sueldos empres ACE'   # define título principal
titx<-'Sueldos (miles de pesos)'   # define título eje horizontal
tity<-'Frecuencia'   # define título eje vertical
hist(sueldos,   # construir histograma
main=tit,
xlab=titx,
lwd=2,   # ancho de linea
ylab=tity,
col='green')   # color de relleno

boxplot(sueldos,   # gráfico de cajón
main=tit,
xlab=titx,
lwd=2,
col='green',
horizontal=T)   # horizontal
```

Los años de escolaridad y los tiempos de viaje se hacen de forma análoga.

EJEMPLO 24

Estos datos se ingresarán como tabla de datos (data frame):

```
funcionarios<-read.table(header=TRUE,    # header=TRUE indica que tienen encabezado
text='   # indica el inicio de los datos
Funcionario   Experiencia   Sueldo   Nivel
1    1     400     2
2    7    1951     1
3    3     952     1
 . . .
15  10   1615     2
')   # fin del ingreso de datos
Str(funcionarios)   # muestra la composición y las características de la tabla de datos
attach(funcionarios)   # permite utilizar las variables en forma independiente de la tabla
color<-Estamento+1   # define colres verde (1) y rojo (2)
rangox<-c(0,20)   # define límites eje horizontal
rangoy<-c(300,3000)   #define límites eje vertical

plot(Experiencia,Sueldo,   # diagrama de dispersión
col=color,   # color puntos
xlim=rangox,   # rango horizontal
ylim=rangoy,   # rango vertical
xlab='Experiencia  (años)',  # título horixontal
ylab='Sueldo  (miles de pesos)',  # título vertical
main='Sueldo quad versus quad Experiencia' ,  # título principal
pch=16,   # tipo de símbolo: círculo
```

```
cex=2)   # tamaño símbolo
text(Experiencia+1,   # rótulos en cada punto: posición horizontal
Sueldo-20,   # posición vertical
labels=Funcionario)   # rótulos
legend("bottomright",   # leyenda en posición abajo a la derecha
title="Estamento")   # título leyenda
title.col="black",   # color título leyenda
legend=c("Administrativo","Profesional"),   # nombres leyenda
text.col=color,   # colores rótulos leyenda
```

EJEMPLO 25

Se usó R para el cuarto y quinto gráficos. Los primeros tres se hicieron con Excel.

```
energia<-read.table(header=TRUE,text='   # ingreso datos
mes   ter      hid      tri   sem
Ene    1475    2480    1     1
....
')    # fin del ingreso de datos
attach(energia)    # permite usar las variables independientes de la tabla de datos
color<-tri+1   # define colores
rango<-c(1400,2800)   # define rango de ambos ejes
plot(ter,hid,col=color,   # diagrama de disperión
xlim=rango,ylim=rango,   # rango de los ejes
xlab=  textit'Térmica (MMKw)',  # rótulo eje horizontal
ylab=  textit'Hidráulica (MMKw)',  # rótulo eje vertical
main=  textit'Generación de energía en 2007',  # título principal
pch=16,   # define tipo de representación: círculos
```

```
cex=2)   # tamaño círculos

text(ter+80,   # pone rótulos en los círculos

hid+10,   # posición horizontal de los rótulos

labels=mes)   # posición vertical


ms<-1:12   # define meses de 1 a 12

plot(ms,ter+hid,     # diagrama de dispersión de mes versus energía total generada

type="o",   # tipo  "o" : une los puntos, representados por círculos

ylim=c(0,4200),   # rango eje vertical

lwd=2,   # ancho de línea

xaxt="n",   #  anular la escritura en el eje horizontal para ser reemplazada por lo definido en  "axis"

col="violet",   # color de las líneas

main=  textit"Generación eléctrica en 2007",  # título principal

xlab=  textit"Mes",  # rótulo eje horizontal

ylab=  textit"Generacion (MMKw)"  # rótulo eje vertical


axis(1,   # define escritura de puntos en el eje horizontal, anulada por  xaxt="n"

at = 1:12,   # determina dónde se escribirán

labels = ms)   # determina qué se escribirá


lines(ms,ter    # para sobreponer otra curva en el mismo gráfico, de energía térmica

type="o",   # tipo de curva

col="red",   # color de la curva

lwd=2)   # ancho de linea
```

```
lines(ms, hid,type="o",col="blue",lwd=2)    # agrega otra curva de energía hidráulica

legend ("bottomleft",  # leyenda en posición abajo izquierda
title= "Generación",  # título de la leyenda
title.col="black",   # color del título
legend= c("Térmica","Hidráulica","Total"),  # rótulos de la leyenda
text.col=c("red","blue","violet"))   # colores de los rótulos
```

La Unidad de Currículum y Evaluación (UCE) del Ministerio de Educación de Chile, desarrolla las definiciones curriculares y evaluativas, que describen los aprendizajes a ser alcanzados por los estudiantes en toda su trayectoria educativa, y de proveer y resguardar la coherencia y alineamiento de los recursos educativos de apoyo para la implementación del currículum nacional.

El propósito formativo de la asignatura de Matemática es enriquecer la comprensión de la realidad, facilitar la selección de estrategias para resolver problemas y contribuir al desarrollo del pensamiento crítico y autónomo en todos los estudiantes, sean cuales sean sus opciones de vida y sean cuales sean sus opciones de estudios al final de la experiencia escolar. La matemática proporciona herramientas conceptuales para analizar la información cuantitativa presente en noticias, opiniones, publicidad y diversos textos, aportando al desarrollo de las capacidades de comunicaciones, razonamiento y abstracción e impulsando el desarrollo del pensamiento intuitivo y la reflexión sistemática.

El Eje Datos y Probabilidades responde a la necesidad de que todos los estudiantes registren, clasifiquen, lean información dispuesta en tablas y gráficos, y que se inicien en temas relacionados con las probabilidades. En este sentido, este libro es un aporte para la implementación de los objetivos de aprendizaje del currículum escolar de Matemática. Es una oportunidad para que docentes y estudiantes puedan acceder al conocimiento que posibilite procesar información proveniente de la realidad y así profundizar su comprensión acerca de ella y de los conceptos aprendidos.