ReFIP Matemática

RECURSOS PARA LA FORMACIÓN INICIAL DE PROFESORES DE EDUCACIÓN BÁSICA

Datos y azar

PARA FUTUROS PROFESORES DE EDUCACIÓN BÁSICA

ReFIP Matemática

RECURSOS PARA LA FORMACIÓN INICIAL
DE PROFESORES DE EDUCACIÓN BÁSICA

Datos y Azar

PARA FUTUROS PROFESORES DE EDUCACIÓN BÁSICA

AUTORES:

Ana María Araneda,
Pontificia Universidad Católica de Chile

Eugenio Chandía,
Universidad de Chile

María Alejandra Sorto,
Texas State University

Proyecto FONDEF – CONICYT D09 I1023 (2011 – 2014)
Directora de Proyecto: Salomé Martínez
Autoría: Ana María Araneda
Eugenio Chandía
María Alejandra Sorto

Registro de propiedad intelectual:
ISBN: 978-956-349-617-8
Depósito legal: 236738

Dirección editorial: Arlette Sandoval Espinoza
Corrección de estilo: María Paz Contreras Aguirre
Dirección de arte: Carmen Gloria Robles Sepúlveda
Coordinación diseño: Vinka Guzmán Tacla
Diseño Portada: José Luis Jorquera Dölz
Diagramación: Ximena Moncada Lomeña
Ilustración: Carlos Valentino Romero Cáceres
Producción: Andrea Carrasco Zavala

Primera edición: diciembre 2013

Coyancura 2283, oficina 2013,
Providencia. Santiago de Chile.

www.ediciones-sm.cl
Atención al cliente: 600 381 13 12

Impreso en Chile/ Printed in Chile

ReFIP Matemática

Recursos para la Formación Inicial de Profesores de Educación Básica en Matemática

Proyecto FONDEF - CONICYT D09 I1023 (2011 - 2014)

Directora:	Salomé Martínez, Centro de Modelamiento Matemático, Universidad de Chile
Director alterno:	Héctor Ramírez, Centro de Modelamiento Matemático, Universidad de Chile
Institución beneficiaria principal:	Centro de Modelamiento Matemático, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile
Institución beneficiaria asociada:	Facultad de Matemáticas, Pontificia Universidad Católica de Chile
Instituciones asociadas:	Ediciones SM Chile Ministerio de Educación Fundación Luksic Academia Chilena de Ciencias

PRESENTACIÓN

Cuando se incuba un sueño y nace una idea, el mundo se ensancha,
se ponen en juego los recursos y se inicia un camino donde los obstáculos
se presentan uno a uno. Cuando la idea es buena y el sueño es grande,
nada detiene el ímpetu desatado.

La colección de libros que tengo el honor de presentar es el fruto de un sueño grande y del trabajo persistente de un equipo formado por matemáticos y educadores matemáticos, quienes, liderados por su directora, se abocaron a la tarea de escribir cuatro libros de matemática, sobre los temas centrales del currículo escolar: Números, Geometría, Álgebra y Datos y Azar.

Estos cuatro libros representan la culminación de un proceso de aprendizaje, reflexión y maduración que se inicia muy atrás, con las primeras iniciativas en educación en el Centro de Modelamiento Matemático, cuando se vislumbraba que era posible hacer un aporte a la educación desde la perspectiva de los matemáticos, pero no se sabía muy bien cómo. Fueron muchos los proyectos que se sucedieron y que fueron ayudando a comprender mejor el problema que se enfrentaba y ayudaron a apuntar con mayor precisión a una de nuestras principales debilidades en matemática escolar: las escasas oportunidades que nuestro sistema de formación de profesores brinda a los estudiantes de Pedagogía en Educación Básica de conocer la matemática escolar. Esta colección de libros apunta, con una potente fuerza de saber, al corazón del sistema formativo, proveyendo matemática en sus contenidos y la manera de enseñarlos.

Si estos libros representan la culminación de un proceso, también son solo un hito en el camino que se abre hacia el futuro con inmensos desafíos, algunos de los cuales son desatados por estos mismos. La incorporación en las aulas universitarias de la matemática escolar, con toda la potencialidad y riqueza que estos libros proponen, requiere de grandes esfuerzos de parte de las propias unidades formadoras, de los formadores de profesores y ciertamente de los estudiantes de pedagogía que sueñan con un aula escolar viva y ávida de conocimiento. Estos libros ponen de manifiesto la necesidad de formación académica de los formadores de profesores y llaman a la creación de material de apoyo complementario en otros formatos. Con una mirada de largo plazo, estos libros también muestran la necesidad de contar con académicos de alto nivel, conocedores de la matemática escolar, de su enseñanza y de su aprendizaje, en todas las unidades de formación de profesores.

El proceso que da vida a esta colección de libros, desde su concepción hasta la impresión final de sus páginas, tiene numerosos rasgos originales que quisiera destacar. Este es un proyecto que convoca a matemáticos interesados por la educación, expresando una realidad creciente en todo el mundo y también en nuestro país, que

mueve a científicos investigadores de sus propias disciplinas a abordar problemas de la educación, con espíritu abierto y con el respeto que merecen. Expresiones de esta tendencia van en la línea de un cambio de parte de los científicos, que han ido comprendiendo la complejidad de los problemas de la educación, de la formación de profesores, de la escuela y la sala de clases. Pero este proyecto también convoca a educadores matemáticos que, por la naturaleza de su disciplina científica, tienen a la educación en toda su complejidad en el centro de su quehacer, pero que en muchas ocasiones han caminado por una vía paralela a los científicos. El proyecto que da origen a los libros que aquí presento es una muestra más de la importancia de acercar estos mundos y de la tendencia nacional a comprender que en la educación hay espacio para todos, que la incorporación de actores enriquece la discusión y mejora la calidad de los resultados. Estos libros son el fruto del trabajo conjunto de matemáticos y educadores matemáticos.

Estos libros no nacen del trabajo aislado de los expertos convocados, sino que en todo momento se ha tenido presente la realidad, expresada a través de la opinión de los actores que intervienen en la formación de los profesores de educación básica. Las necesidades, el sentir y las opiniones de los académicos formadores y de los estudiantes de pedagogía fueron recogidos en consultas y aplicaciones piloto a lo largo de todo el país. Esta es una experiencia inédita en Chile, que incorpora a los lectores en la redacción de libros de texto universitarios, basando las decisiones editoriales en la evidencia encontrada, sobre lo que es relevante para el profesor, y dando fuerza a las ideas que se presentan en sus páginas. Es interesante que en la búsqueda de información para apoyar la escritura de los libros, los autores tuvieron la oportunidad de dar una mirada nacional a la formación de los profesores de educación básica en cuanto a la matemática, la que les permitió levantar evidencia de investigación que resulta de extremo interés, más allá de su propósito original.

Estos libros sobre la matemática escolar son una herramienta poderosa para apoyar la formación de profesores, ya que enfocan los contenidos matemáticos conectados con su enseñanza y teniendo en cuenta el currículo nacional. Así como sus cuatro tomos van tomando uno a uno los temas centrales del currículo, con una mirada puesta en los contenidos escolares, proyectados en la sala de clases. Dan cuenta de la matemática escolar en todas sus dimensiones, las que muchas veces son minimizadas equivocadamente ignorando su complejidad. Una lectura de sus páginas nos lleva a comprender rápidamente que la tarea de enseñar matemática escolar es intelectualmente demandante y que requiere de una cuidadosa preparación, que va mucho más allá de unos cursos aislados. Estos libros muestran la importancia de la comprensión de los contenidos, teniendo presentes las diversas formas de enseñanza y de aprendizaje y el currículo escolar, y sugieren un cambio importante en el eje de las carreras de pedagogía, moviéndolo desde una mirada generalista desprovista de contenido a una mirada integradora del contenidos y su enseñanza.

En la línea de esta última reflexión, con la publicación de esta obra se plantean en forma concreta lineamientos respecto de cómo deberíamos formar a los profesores en Chile. Debemos transformar una cultura universitaria que considera que la disciplina

que se enseña es secundaria frente a un saber pedagógico general y teórico, desde el cual sería posible deducir qué hay que hacer en el caso de cada disciplina. Debemos transformar una cultura universitaria que considera que en la formación de profesores, la preparación disciplinaria y pedagógica van en paralelo, dejando que el estudiante de pedagogía haga la integración. Es necesario movernos a una cultura de la integración entre las disciplinas y lo pedagógico, generando un compromiso de los formadores que abordan estos aspectos en forma coordinada e integrada. Ciertamente la formación de un profesor va más allá de lo disciplinario y lo pedagógico, pero si estos aspectos no están presentes con fuerza y en forma integrada, no tendremos un profesor o profesora con la potencialidad de proyectar el saber formador en su integridad.

Estos libros nacen en un contexto marcado por una creciente preocupación nacional por la educación, empujada por las demandas del movimiento estudiantil en sus múltiples expresiones. Esta preocupación pone énfasis en el acceso y la calidad de la educación, y demanda importantes recursos para la introducción de los cambios estructurales necesarios, que garanticen el acceso de todos los niños y niñas a la educación de calidad. Sin embargo, es necesario hacer notar que con una inyección importante de recursos y con una juiciosa reorganización administrativa, la calidad de la educación no queda garantizada. En este contexto, es importante mencionar que la colección de libros que presentamos nace en el seno del programa INICIA, lanzado en 2008 y que tiene como propósito el fortalecimiento de la formación inicial de los profesores. Estos libros nacen y se nutren de las experiencias adquiridas en la formulación de los estándares de matemática, que definen lo que como país esperamos que los futuros egresados de las carreras de pedagogía sepan y sepan hacer, y que se miden en la prueba INICIA. Los estándares y la prueba INICIA definen un marco de demandas para los formadores de profesores y para los estudiantes de pedagogía difíciles de lograr sin tener el apoyo del Estado, de las instituciones y de académicos preocupados por el avance de la calidad de la educación. Esta colección ofrece apoyo en el área de matemática e interpreta los estándares desde una perspectiva de la realidad nacional. Bienvenidos serán materiales complementarios que ayuden a formar mejores profesores y profesoras.

Me sumo con alegría a todos los que ven en estos libros una poderosa herramienta para seguir construyendo una mejor educación para todos nuestros niños y niñas, y a todos quienes levantan su voz para felicitar a la directora y a todos los autores de estos libros, quienes con su trabajo, talento y perseverancia nos ponen un desafío más en esta tarea de hacer de Chile un país con una educación justa y de calidad, donde todos los niños y niñas tengan acceso a una educación que les permita conocer las matemáticas, las ciencias, las humanidades, las artes y todas las expresiones de la cultura humana, para construir así un país mejor, un país desarrollado.

Patricio Felmer
Premio Nacional de Ciencias Exactas 2011
Académico de la Universidad de Chile

AGRADECIMIENTOS

La colección de textos ReFIP fue desarrollada como parte del proyecto FONDEF D09I1023 "Recursos pedagógicos para la implementación de los Estándares de Formación Inicial de profesores de Educación Básica en matemática" (ReFIP). Agradecemos el apoyo de Conicyt a través del programa FONDEF, el cual fue clave para la realización de esta colección. En particular, reconocemos el apoyo del Comité de Área de Educación de FONDEF, y muy especialmente la dedicación de Daniela Fuentes, ejecutiva a cargo del proyecto.

Expresamos también nuestra gratitud a la institución que albergó a este proyecto, el Centro de Modelamiento Matemático de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile, en particular, al director del CMM, Alejandro Jofré, y a Francisco Brieva, decano de la Facultad de Ciencias Físicas y Matemáticas, por proveernos el soporte que este proyecto necesitó. Asimismo, agradecemos todo el apoyo de la Facultad de Matemáticas de la Pontificia Universidad Católica de Chile, y muy especialmente a su decano, Martin Chuaqui.

Las instituciones asociadas al proyecto han sido esenciales para su desarrollo, en particular, ellas apoyaron con decisión su presentación. Agradecemos a Ediciones SM, en particular a su gerente general, Francisco Tepper, y a su directora editorial, Arlette Sandoval. También queremos agradecer el patrocinio de Fundación Luksic, en especial a Monserrat Baranda, su gerente general. Agradecemos también a la directora del Centro de Perfeccionamiento e Investigaciones Pedagógicas del Ministerio de Educación (CPEIP), Paula Pinedo, y a Regina Silva, Coordinadora del Área de Educación Continua (CPEIP). Todo nuestro reconocimiento va también a la Academia Chilena de Ciencias y a su presidente, Juan Asenjo, por la permanente colaboración.

Un hito importante en el desarrollo de los textos fue la utilización de sus versiones preliminares en cursos de matemáticas de carreras de Pedagogía en Educación Básica. Estos pilotos se desarrollaron en 16 universidades: Pontificia Universidad Católica de Chile, Universidad Alberto Hurtado, Universidad Arturo Prat, Universidad Católica de la Santísima Concepción, Universidad Católica de Temuco, Universidad de Concepción, Universidad de las Américas, Universidad del Bío-Bío, Universidad del Desarrollo, Universidad de Los Andes, Universidad de Magallanes, Universidad de Playa Ancha, Universidad de Viña del Mar, Universidad Diego Portales, Universidad Santo Tomás, Universidad San Sebastián y Universidad de los Andes. Estos pilotos no podrían haberse llevado a cabo sin el apoyo de las autoridades de estas universidades, quienes tuvieron una confianza enorme en nuestro equipo y nos apoyaron en todas las actividades de esta etapa. Agradecemos especialmente a sus académicos formadores y a los estudiantes de los cursos donde se probaron nuestros textos. Valoramos su generosidad y activa participación en las distintas actividades del proyecto.

Durante el desarrollo del proyecto contamos con la guía del Comité Asesor, conformado por Patricio Felmer, Miguel Díaz, Raimundo Olfos, María Aravena y Arturo Mena, quienes evaluaron versiones preliminares de los textos y orientaron nuestro trabajo. Valoramos sus aportes a lo largo del proyecto. En esta misma línea, agradecemos a Pablo Dartnell por sus valiosas contribuciones.

Agradecemos también a los estudiantes, académicos formadores y evaluadores nacionales e internacionales que nos entregaron sugerencias y comentarios, que ayudaron a enriquecer la colección de textos. También reconocemos el valioso trabajo del equipo editorial de Ediciones SM en la etapa final de producción de los textos.

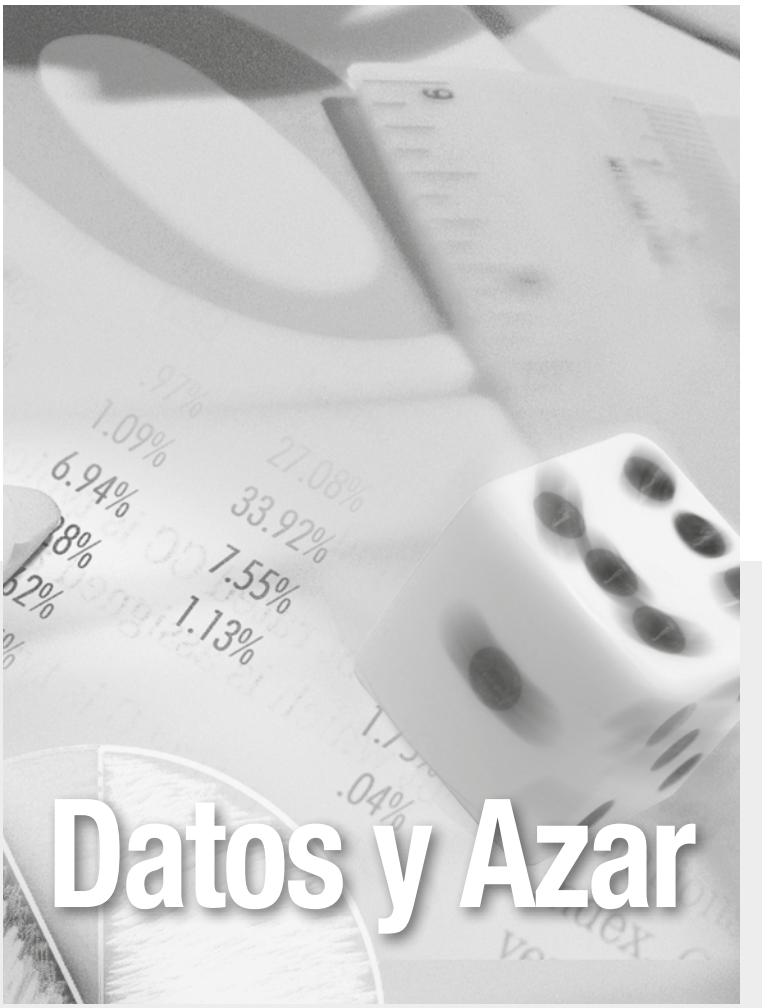
Queremos agradecer todo el apoyo en la gestión y administración del proyecto. Agradecemos a Erika Pino, Paulina Zavala y María Eugenia Heckmann de la Pontificia Universidad Católica de Chile, a Judith Figueroa, Eterin Jaña y Silvia Mariano de la Universidad de Chile, y muy especialmente a María Cecilia Cea de la Universidad de Chile. También agradecemos a Bárbara Salas por el apoyo en la difusión del proyecto.

Valoramos también la disposición de Carmen Montecinos y José Sánchez a ser parte del Comité Editorial de esta colección.

Finalmente, queremos expresar nuestra gratitud a dos connotados académicos que influyeron fuertemente en nuestro quehacer. A Sybilla Beckmann, académica de la Universidad de Georgia y autora de un reconocido libro de matemática para la formación de profesores en Estados Unidos, por sus valiosos consejos que fueron una guía durante todo el proyecto, y a Patricio Felmer, académico de la Universidad de Chile y referente nacional en temas relacionados con educación matemática, por su generosa ayuda.

Salomé Martínez
Directora del Proyecto
Fondef D09I1023

Héctor Ramírez
Director Alterno del Proyecto
Fondef D09I1023



Datos y Azar

ÍNDICE

INTRODUCCIÓN

Cada día, todo ciudadano se ve enfrentado a información construida en base a datos. Esta información, utilizada de manera adecuada, le servirá de fuente tanto para entender los fenómenos que lo rodean, como para tomar decisiones de manera informada. Ya sea en el papel de investigador, analizando datos y transformándolos en información, o como receptor de esta, todo ciudadano debe estar preparado para aprovechar esta oportunidad. Estar capacitados para hacer buenas preguntas, usar datos en forma inteligente, evaluar conjeturas basadas en ellos y formular conclusiones son habilidades básicas en la sociedad actual.

Acorde a esto, desde la Reforma Educacional de los años noventa en Chile, los contenidos relacionados a estadística y probabilidad han tenido cabida en el currículo escolar. A partir del Ajuste Curricular del año 2009, estos contenidos aparecen desde primero básico hasta cuarto medio, como uno de los contenidos conductores de toda la enseñanza escolar obligatoria chilena en Matemática, en el eje de Datos y Probabilidades. Concordantemente, estos contenidos han pasado a ser parte de los Estándares Orientadores para egresados de las carreras de Pedagogía en Educación Básica [1] en el área de Matemática del año 2011.

Todo profesor, como actor fundamental en la experiencia educativa de sus alumnos, debe estar capacitado para crear constantemente oportunidades para que los alumnos adquieran las nuevas habilidades requeridas. Es esperable que el profesor involucre a sus alumnos desde muy temprana edad en el manejo directo de datos, y que este proceso crezca en sofisticación y complejidad a medida que los alumnos progresan en sus estudios.

Inspirados en el informe GAISE [2] podemos decir que la expectativa sobre los alumnos es que, a la edad apropiada, aprendan a: formular preguntas pertinentes y hacer conjeturas a partir de datos o situaciones en las que interviene el azar y clasificar, organizar, resumir y representar datos y analizarlos críticamente para obtener información a partir de estos. En el área de la probabilidad, la expectativa sobre los alumnos es que ellos comprendan y sean capaces de usar el lenguaje de probabilidades, determinen la probabilidad de ocurrencia de eventos en forma experimental y teórica a partir de fenómenos aleatorios, y analicen resultados. Se espera que estas habilidades cimienten las bases para el estudio formal de la probabilidad durante la Educación Media.

Lo anterior muestra el enorme desafío que implica capacitar a futuros profesores en la tarea que deben desempeñar. Ante esta situación, el presente libro tiene como propósito desarrollar en ellos la habilidad de introducir ideas básicas de estadística y probabilidad, en forma deliberada y con propósito, en los primeros años, para así ahondar y expandir el entendimiento de sus alumnos. El libro vincula de forma explícita los conceptos de un curso de

[1] Recuperado el 1 de marzo del 2013 en: http://www.cpeip.cl/usuarios/cpeip/File/2012/librobasicaokdos.pdf

[2] Recuperado el 1 de marzo del 2013 en: http://www.amstat.org/education/gaise/GAISEPreK12_Intro.pdf

estadística y probabilidad para futuros profesores a nivel universitario, con los conceptos que profesores de Educación Básica deben enseñar a sus alumnos. Para alcanzar este propósito, se presentan los contenidos de manera informal –sin sacrificar la integridad de la disciplina– y se ilustran con ejemplos de situaciones de aula, actividades y ejercicios.

Cada capítulo está organizado para que el futuro profesor aprenda y ejercite los conceptos importantes, y realice interpretaciones que faciliten su aprendizaje. Cada capítulo, además, motiva la necesidad de contar con nuevas herramientas, presentando luego los contenidos disciplinares asociados para suplir dicha necesidad. Por último, se plantea una discusión sobre errores frecuentes de los alumnos relacionados a los tópicos propios de los capítulos.

En la esencia de este texto, cada capítulo motiva la reflexión a través del análisis de situaciones cotidianas y del planteamiento frecuente de inquietudes destinadas a ayudar a los alumnos de Pedagogía en Educación Básica en el desarrollo de su propio conocimiento.

Esta esencia no se hubiera alcanzado sin la ayuda de quienes compartieron con nosotros sus enfoques, comentarios y sugerencias. Agradecemos: a Guido del Pino por compartir con nosotros su visión de la estadística y el rol de la probabilidad, y por sugerirnos una forma de presentar los contenidos de probabilidad evitando notación y herramientas innecesarias; a Nancy Lacourly y Servet Martínez, por su lectura detenida del libro y cuyos comentarios y aportes apreciamos enormemente; y a María José García, por su importante aporte en la creación del Capítulo III, y por su lectura minuciosa de versiones finales.

Anita Araneda agradece a su familia: a Álvaro, Javiera y Josefa por su apoyo y por la generosidad con que compartieron el tiempo dedicado a este libro; a Matías, Nicolás y Joaquín por su cariño constante. Eugenio Chandía agradece a Roxana, Katherine y Estefany por comprender, darle la energía, la fuerza y el tiempo cuando lo necesitaba, sin ustedes no lo hubiera logrado. También agradece a Salomé, por darle la oportunidad de trabajar en este hermoso proyecto y de poder trabajar con Anita y Alejandra; fue realmente muy provechoso para él trabajar con estas dos destacadas profesoras en el área de la Educación Estadística. María Alejandra Sorto da las gracias a sus hijas Isabel y Sofía, y a su marido Alexander White por las innumerables conversaciones y valiosas sugerencias acerca de la presentación de conceptos estadísticos.

Capítulo
I

Ciclo de investigación

Introducción

El *ciclo de investigación*, íntimamente ligado al denominado método científico, constituye un procedimiento empírico-analítico riguroso que, a través del tiempo, ha permitido numerosos avances en la generación de conocimiento en las ciencias. Galileo Galilei es considerado el creador del método científico, a través del cual postuló la ley de gravedad, que fue posteriormente formulada por Isaac Newton. El descubrimiento de la penicilina, por Alexander Fleming en el siglo XX, corresponde a otro ícono de los avances permitidos por este método.

En este capítulo, estudiaremos el ciclo de investigación, por una parte, como un proceso científico para responder preguntas de interés y, por otra, como un método de enseñanza y aprendizaje de la estadística escolar. Este proceso, que comienza con la observación de fenómenos y la formulación de hipótesis o preguntas relacionadas a estos, promueve la argumentación en base al análisis de información extraída de datos que provienen de experimentos o simulaciones, creando así nuevo conocimiento respecto al fenómeno de interés. Por lo tanto, se ha escogido este método en la enseñanza de la estadística, ya que permite integrar tanto los conocimientos propios de esta disciplina, como los relacionados con la ciencia.

El ciclo de investigación permite que niños en edad escolar comprendan la importancia de las herramientas estadísticas en la búsqueda de nuevo conocimiento. Estas facilitan la obtención de respuestas a preguntas acerca de fenómenos que ellos mismos observan, adquiriendo conocimiento respecto de estos y generando luego nuevas interrogantes que motivan a repetir el proceso, convirtiéndolo en un ciclo de aprendizaje continuo. Por otra parte, el hecho de abordar la enseñanza de la estadística a través del ciclo de investigación promueve una actitud positiva de los alumnos hacia las ciencias. Los profesores en Educación Básica disponen continuamente de oportunidades para nutrir la curiosidad innata de sus alumnos, y demostrarles que ellos mismos pueden dar respuesta a sus propias preguntas mediante la búsqueda y recolección de información.

Este capítulo se organiza como sigue: la **Sección 1** trata cada una de las etapas que forman el ciclo de investigación y describe algunas formas de abordarlas en el aula escolar. La **Sección 2** detalla la presencia de las etapas del ciclo de investigación en el actual currículo chileno de Educación Básica, y discute una actividad de aula escolar, en el contexto del ciclo de investigación.

1. Etapas del ciclo de investigación

El ciclo de investigación considera cinco etapas: *Planteamiento del problema*, *Planificación*, *Recolección de datos*, *Análisis* y obtención de *Conclusiones*, las que se muestran en la **Figura I.1**. Al finalizar un ciclo, es esperable que el investigador haya logrado cierto aprendizaje con respecto a la pregunta de interés que se ha planteado en la primera etapa. Las etapas que se muestran en la figura constituyen un ciclo, en el sentido de que la adquisición de nuevo conocimiento debiese generar nuevas preguntas de interés para el investigador, donde la obtención de respuestas requerirá de la repetición de las etapas.

En los siguientes apartados, estudiaremos cada una de las etapas mencionadas.

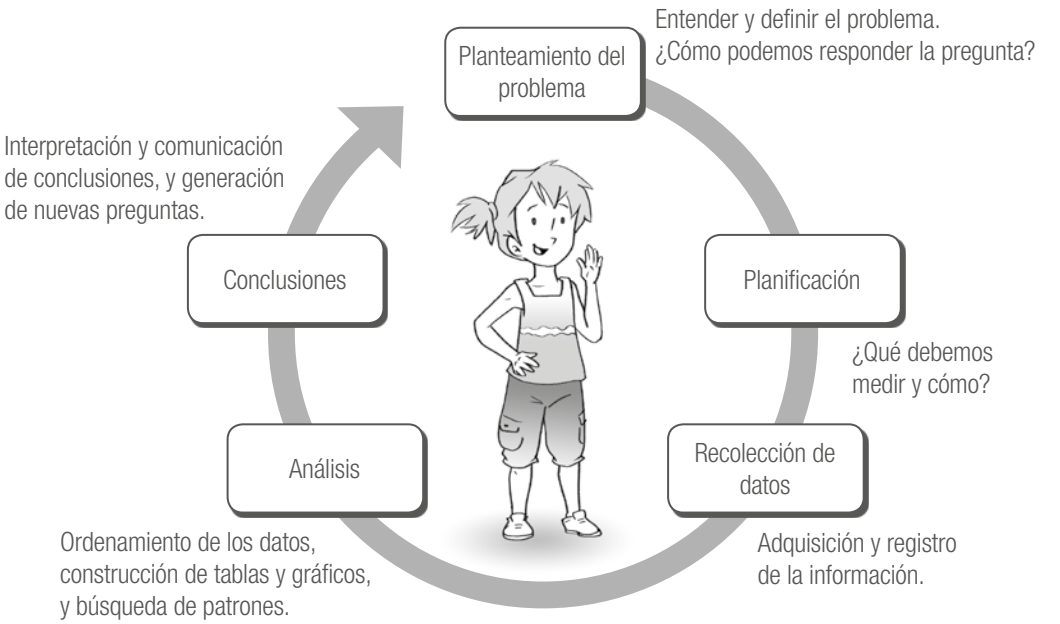


Figura I.1: Etapas del ciclo de investigación.

1.1. Planteamiento del problema, planificación y recolección de datos

En la primera etapa del ciclo de investigación, *Planteamiento del problema*, el investigador debe identificar el fenómeno sobre el cual desea aprender, formalizando la o las preguntas que quiere responder a través del proceso. En ocasiones, esto también incluye formular hipótesis que serán aceptadas o rechazadas en función de la evidencia obtenida. A modo de ejemplo, en algunas investigaciones médicas, equipos de profesionales están interesados en estudiar la efectividad de una nueva droga en el tratamiento de una enfermedad. En estos casos, la hipótesis suele ser que esta droga es más efectiva que las drogas o tratamientos alternativos habituales.

A nivel escolar, en los primeros niveles, los niños están primordialmente interesados en fenómenos relacionados con ellos mismos y sus alrededores, y sus preguntas se centran en temas como ¿cuántos compañeros y compañeras del curso caminan a la escuela? o ¿cuál es el equipo de fútbol favorito del curso? A medida que los alumnos progresan en su educación, sus intereses se van expandiendo a su comunidad o país. Sus preguntas, entonces, pueden ser más complejas y

sofisticadas, como, por ejemplo, ¿qué tan interesada está mi comunidad en reciclar sus desechos? o ¿existe relación entre el sexo y los intereses deportivos de niños de mi edad? Cuando esto sucede, la cantidad de datos necesarios para responder a sus preguntas crece y la tarea de recolección y organización gana en complejidad.

La necesidad de responder estas preguntas requerirá del uso de herramientas de recolección y análisis de datos, por lo que a este tipo de preguntas se les denomina *preguntas estadísticas*.

Luego de definir la o las preguntas de interés, es necesario determinar la forma de responderlas. Para esto se deberá precisar qué información o datos es necesario recolectar, de qué forma se hará, y tomar decisiones sobre cómo registrarlos, entre otras consideraciones. Algunas veces, los niños y niñas recolectarán datos por sí mismos; otras veces, harán uso de bases de datos ya existentes en distintas fuentes. La determinación de todos estos aspectos mencionados pertenece a la etapa de *Planificación*. Una vez definidos los aspectos descritos, se debe poner en marcha la etapa de *Recolección de datos*.

En estas tres etapas, se debe procurar que los alumnos aprendan cuidadosamente a enmarcar sus preguntas y a determinar qué datos recolectar, y cuándo y cómo recolectarlos. Ellos aprenderán a reconocer diferencias entre variadas formas y técnicas para recolectar datos, incluyendo la observación, medición, experimentación, y estudios a través de encuestas, así como también notarán cómo la forma en que la pregunta está planteada ayuda a determinar la manera más apropiada de recolectar la información.

En los primeros años de escolaridad, las preguntas de investigación formuladas usualmente se limitan al aula de clases o a otros grupos pequeños. Esto permite que la recolección de datos corresponda, en realidad, a un *censo*. Al ir madurando, los alumnos comienzan a comprender que la razón principal de recolectar y analizar datos es hacer inferencias y predicciones que puedan ser aplicadas solo a un conjunto de datos dado.

A medida que se avanza en la Educación Básica, es necesario que los alumnos se den cuenta de la importancia de la articulación de una buena pregunta, y de la planificación cuidadosa de la manera de obtener la información y su posterior representación. Así descubrirán que organizar y ordenar datos les ayudará a encontrar las respuestas que buscan. Sin embargo, aprender a refinar sus preguntas y planificar en forma efectiva la recolección de datos son habilidades que los alumnos solo desarrollan mediante mucha repetición de experiencias, discusiones frecuentes y la hábil guía de sus profesores. Para alcanzar esto, es necesario que en el aula escolar se den oportunidades para que los alumnos planteen preguntas interesantes y desarrollen maneras de recolectar datos que les ayudarán a formular respuestas.

1.2. Análisis y conclusiones

Una vez recolectados los datos, es necesario organizarlos, clasificarlos y ordenarlos, para encontrar patrones y regularidades. Aquí, es posible construir o completar tablas y gráficos, o determinar representantes de comportamiento, como medidas de tendencia central y medidas de dispersión, para poder analizar la distribución de las características o atributos en el conjunto de datos. A esta etapa se le denomina *Análisis*.

Los profesores deben estimular la reflexión sobre el problema a partir del despliegue y representación de los datos que sus alumnos hayan recolectado, aun cuando ellos se encuentren en niveles escolares iniciales. Preguntas como ¿hay más niños en nuestro curso que prefieren el helado de frutilla, o hay más niños que prefieren el helado de manjar? incitan a los alumnos a escudriñar en los datos en la búsqueda de información, lo que conlleva la necesidad de que dichos datos estén organizados y ordenados. Al contestar preguntas como la anterior, los alumnos también se dan cuenta de que la información obtenida a partir de los datos recopilados puede ayudar a tomar decisiones, tales como qué cantidad de helado de frutilla y de manjar encargar para una celebración.

Inicialmente, los alumnos pueden utilizar objetos concretos, o sus representaciones, para dar respuesta a sus preguntas, como papeles lustres para representar el color de ojos de cada uno de sus compañeros. Estos objetos pueden ser organizados y representados en un gráfico concreto, de tal manera que se pueda responder alguna pregunta de interés. Los niños deben ir avanzando en la abstracción de la representación utilizada. En efecto, en Educación Básica se abordan luego otras representaciones, como tablas, gráficos de puntos, gráficos de barras, gráficos circulares y gráficos de tendencia, cuyo análisis requiere que los alumnos desarrollen habilidades de lectura e interpretación. Al examinar, comparar y discutir muchos ejemplos de conjuntos de datos y sus representaciones, los alumnos ganan un entendimiento importante acerca de las diferencias entre tipos de datos –de carácter cualitativo o cuantitativo–, la necesidad de seleccionar las escalas apropiadas para los ejes de los gráficos y las ventajas de las diferentes representaciones gráficas para destacar distintos aspectos de un mismo conjunto de datos.

En los niveles superiores de Educación Básica se espera que los niños examinen la distribución de datos como un todo, así como también que comparen dos o más conjuntos de datos. Para esto, se debe promover que los alumnos examinen aspectos como los valores donde se acumulan los datos, el intervalo de valores de las observaciones, o medidas de tendencia y dispersión. Esto hace necesario que los alumnos adquieran un lenguaje estadístico que les permita comunicar las características de los datos.

En la etapa de *Conclusiones*, los alumnos deben responder la pregunta de interés, así como también deben ser capaces de realizar inferencias y predicciones basadas en el análisis de los datos. Los profesores deben ser activos en promover el desarrollo de las habilidades requeridas para este fin. A modo de ejemplo, tras discutir los resultados de una encuesta para determinar su programa de televisión favorito, un grupo de niños de primero básico podría concluir que un primero básico paralelo de la misma escuela obtendría resultados similares, pero que un grupo de niños de quinto básico obtendría resultados muy diferentes. En este caso, los niños de primero básico podrían especular sobre los motivos que causarían la similitud y la diferencia entre los grupos.

Analizados los datos, se debe estimular a los alumnos a que argumenten, en base al análisis y al conocimiento que se generó a través del ciclo, la respuesta que se dió a la pregunta de investigación, así como también se los debe estimular a realizar conjeturas y nuevas interrogantes acerca del fenómeno observado.

Cuando los niños cursan niveles superiores de Educación Básica, se espera que su destreza para obtener conclusiones, hacer pronósticos y construir argumentos basados en los datos se expanda. A medida que los alumnos ganan experiencia, ellos deben comprender que los datos recopilados en el aula o en su escuela pueden o no ser representativos de una población de alumnos más amplia. Los niños de niveles superiores también deben discutir la diferencia entre distintas

muestras y los factores involucrados en dichos resultados. El planteamiento de hipótesis y el diseño de investigaciones para probar dichas hipótesis pueden ser introducidos en estos niveles.

En resumen

- El *ciclo de investigación* es un proceso científico que permite responder preguntas de interés acerca de un fenómeno.
- Su estudio en Educación Básica promueve la búsqueda de nuevo conocimiento y facilita una actitud positiva de los alumnos hacia las ciencias.

2. El ciclo de investigación en el currículo escolar chileno

Si bien en el currículo escolar se pueden observar todas las etapas del ciclo de investigación, existe un énfasis en la etapa de *Análisis*. A modo de ejemplo, los objetivos de aprendizaje en primero básico corresponden a:[1]

1. *Recolectar y registrar datos para responder preguntas estadísticas sobre sí mismo y el entorno, usando bloques, tablas de conteo y pictogramas.*
2. *Construir, leer e interpretar pictogramas.*

El primer objetivo permite cubrir o hacer alusión a todas las etapas del ciclo de investigación. Sin embargo, el segundo objetivo solo se refiere a la etapa de *Análisis*, a través de la construcción y lectura de pictogramas.

Posteriormente, a modo de ejemplo, los objetivos de aprendizaje en tercero básico corresponden a:

1. *Realizar encuestas, clasificar y organizar los datos obtenidos en tablas y visualizarlos en gráficos de barra.*
2. *Registrar y ordenar datos obtenidos de juegos aleatorios con dados y monedas encontrando el menor, el mayor y estimando el punto medio entre ambos.*
3. *Construir, leer e interpretar pictogramas y gráficos de barra simple con escala, de acuerdo a información recolectada o dada.*
4. *Representar datos usando diagramas de puntos.*

Aunque estos objetivos incluyen algunos aspectos de la etapa de *Recolección de datos*, también se encuentran bastante inclinados hacia la adquisición de conocimientos y destrezas asociadas a la etapa de *Análisis*, ya que menciona la construcción de tablas y variadas representaciones gráficas, así como la obtención de algunos representantes numéricos del conjunto de datos. Algo muy similar ocurre en los niveles escolares básicos restantes.

[1] Tomados de la página web del Ministerio de Educación. http://www.mineduc.cl

A pesar de este aparente y excesivo énfasis dado a la etapa de *Análisis* en el currículo escolar, debemos considerar que, en parte, esta impresión se debe solo a la profundidad en que se describen los objetivos. En efecto, a modo de ejemplo, en los objetivos de aprendizaje de tercero básico mencionados anteriormente, el tercer objetivo no se refiere explícitamente a la etapa de *Recolección de datos*. Sin embargo, el profesor puede tratar esta etapa haciendo mención al proceso de obtención de los datos que se utilizarán en las representaciones. De este modo, la etapa de *Recolección de datos* no necesariamente se encuentra ausente, en favor de la etapa de *Análisis*, sino que los objetivos relacionados a esta última han sido mayormente detallados en la medida que aumenta su complejidad.

En efecto, la **Tabla I.1**, construida a partir de los objetivos de aprendizaje de cada nivel, indica la presencia o ausencia de las etapas del ciclo de investigación en cada uno de los niveles de Educación Básica. En ella, vemos que no solo la etapa de *Análisis* está presente en todo nivel, sino que también están la etapa de *Planificación* y *Recolección de datos*.

	Nivel Escolar					
Etapa del ciclo de investigación	1º básico	2º básico	3º básico	4º básico	5º básico	6º básico
Planteamiento del problema	●	●				●
Planificación	●	●	●	●	●	●
Recolección de datos	●	●	●	●	●	●
Análisis	●	●	●	●	●	●
Conclusiones	●	●				●

Tabla I.1: Presencia de las etapas del ciclo de investigación en cada uno de los niveles de Educación Básica, a partir de los objetivos de aprendizaje.

También se debe tener presente que, aun cuando alguna etapa no se identifique de manera directa en los objetivos de aprendizaje, esta puede observarse de manera indirecta en el currículo escolar, a través de las habilidades declaradas en cada uno de los niveles, que deben ser tratadas de manera a transversal a los objetivos de aprendizaje. En efecto, las habilidades declaradas en el currículo escolar de Educación Básica, corresponden a:

- Resolver problemas
- Argumentar y comunicar
- Modelar
- Representar

De esta manera, a modo de ejemplo, aunque en tercero básico no se observa directamente la etapa de *Planteamiento del problema*, esta puede ser tratada a modo de habilidad. De esta misma forma puede abordarse la etapa de *Conclusiones*, ya que los alumnos deben argumentar y comunicar al responder las preguntas de investigación. Así, el profesor puede transmitir a sus alumnos la integridad del ciclo de investigación en cada nivel de Educación Básica.

Ejemplo:

Aquí se muestra una manera de abordar todas las etapas del ciclo de investigación, a través de una actividad que, aparentemente, solo comprende una parte de este. La siguiente actividad simula ser tomada de un texto escolar.

Es común que en muchas escuelas los niños se organicen para reunir fondos para la Teletón.

En dos meses, los primeros básicos han recolectado $30.000, los segundos básicos $50.000, los terceros básicos $15.000, los cuartos básicos $35.000, y los quintos básicos han recolectado $45.000.

Ordena en la siguiente tabla los datos ordenados por nivel.

Nivel	Monto reunido

Ordena en la siguiente tabla los datos por cantidad reunida (la mayor cantidad primero)

Nivel	Monto reunido

Responde las siguientes preguntas:

1) *¿En cuál de las tablas te es más fácil averiguar qué curso aporta con más dinero?*

2) *¿En cuál de las tablas te es más fácil averiguar la diferencia entre el dinero recolectado por los primeros y los segundos básicos?*

Notamos en la actividad que, si bien existe un contexto, no existe una pregunta de investigación que haya motivado la recolección de datos. En efecto, solo se pide completar ciertas tablas y notar su utilidad. Esta actividad pertenece a la etapa de *Análisis*.

Para abordar la primera etapa del ciclo de investigación, *Planteamiento del problema*, el profesor tiene la posibilidad de enriquecer la actividad estimulando a los niños a plantear preguntas o problemas que pueden haber motivado el análisis que se observa; como, por ejemplo, ¿cuál será el curso que recolectó una mayor cantidad de dinero? o ¿cuál es la diferencia entre el dinero recolectado por los primeros y segundos Básicos?

Luego, para abordar la segunda etapa, *Planificación*, se podrían realizar preguntas a los niños para saber cómo se obtuvieron las cantidades que los cursos recolectaron, lo que requiere determinar, por ejemplo, momentos y maneras de hacerlo. Por último, para motivar la extracción de conclusiones, el profesor debe procurar que los niños respondan las preguntas de investigación anteriormente planteadas.

Como muestra la discusión previa, aunque las actividades de los textos escolares pueden no abordar completamente todas las etapas del ciclo de investigación, los profesores tienen la posibilidad de tratarlas aprovechando la potencialidad de las actividades, utilizando su contexto, los tipos de análisis realizados, el ámbito numérico y los tipos de datos, entre otros.

En resumen

- El currículo escolar permite tratar todas las etapas del ciclo de investigación, a través de los objetivos de aprendizaje y habilidades transversales ahí declaradas.
- El profesor debe aprovechar las potencialidades que entrega este currículo, para que los niños aborden todas las etapas del ciclo de investigación.

Ejercicios

1. Observe los siguientes fenómenos:

   1. El clima de Concepción.
   2. El crecimiento de una planta.
   3. Los medios de transporte en la Región Metropolitana.
   4. La asistencia de niños y niñas a escuelas.
   5. Juegos que niños y niñas realizan en el recreo.

   a. Para cada uno de ellos, escriba una pregunta de interés que pueda ser abordada mediante el ciclo de investigación.
   b. Elabore tres preguntas diferentes para cada uno de los fenómenos, tales que una pueda ser respondida en primero básico, otra en cuarto y la última en sexto, mediante la aplicación del ciclo de investigación.

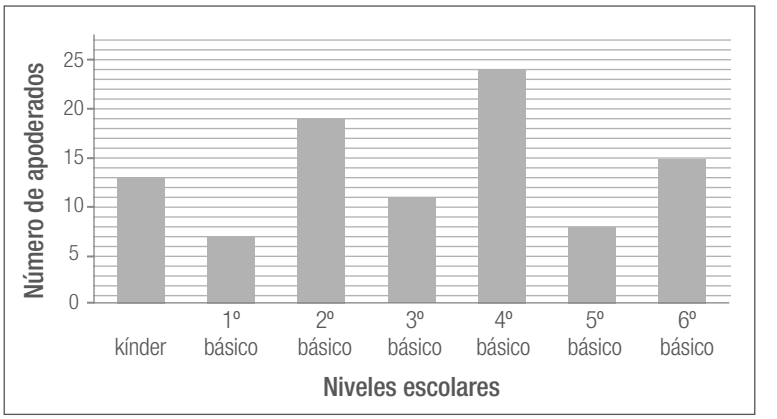
2. Observe los siguientes objetivos de aprendizaje declarados para sexto básico en el eje de Datos y Probabilidades.

   1. Comparar distribuciones de dos grupos provenientes de muestras aleatorias usando diagramas de puntos y de tallo y hojas.
   2. Conjeturar acerca de la tendencia de los resultados obtenidos en repeticiones de un mismo experimento con dados, monedas u otros, de manera manual y/o usando software educativo.
   3. Leer e interpretar gráficos de barra doble y circulares, y comunicar sus conclusiones.

   a. Determine las etapas del ciclo de investigación que podrían abordarse en cada uno de los objetivos de aprendizaje.
   b. Estudie si, a través de las habilidades planteadas de manera transversal, es posible abordar aquellas etapas que no aparecen explícitamente en los objetivos de aprendizaje.

3. La siguiente actividad simula ser tomada de un texto escolar.

Observa el siguiente gráfico, que muestra el número de apoderados que asisten a la reunión de cada curso en el mes de marzo.



1) *¿En qué curso asistieron más apoderados?*

2) *¿En qué curso asistieron exactamente **10** apoderados?*

3) *¿Se puede afirmar que mientras más alto el nivel del curso, mayor es el número de apoderados que asiste?*

a. Determine las etapas del ciclo de investigación que aborda la actividad.

b. Utilizando el contexto de la actividad, elabore instrucciones que permitan tratar las etapas del ciclo de investigación que no son tratadas directamente.

4. Observe el siguiente objetivo de aprendizaje:

Leer, interpretar y completar tablas, gráficos de barra simple y gráficos de línea y comunicar sus conclusiones.

a. Identifique un fenómeno que sea de interés para niños y niñas de quinto básico y cree una actividad que aborde el objetivo de aprendizaje descrito, y las cinco etapas del ciclo de investigación.

b. Identifique las habilidades transversales que permite abordar la actividad.

5. A continuación, se presenta la descripción del Nivel de Aprendizaje Adecuado para el eje de Datos y Probabilidades de los Estándares de Aprendizaje de cuarto básico:

... los estudiantes que alcanzan el Nivel de Aprendizaje Adecuado son capaces de inferir información de acuerdo con datos presentados en tablas, pictogramas y gráficos de barras, y de aplicar dicha información para responder preguntas directas o resolver problemas rutinarios de uno o dos pasos.

a. Identifique la o las etapas del ciclo de investigación que un niño o niña debiera comprender para lograr el nivel que menciona el Estándar de Aprendizaje citado.

b. Observe la descripción del Nivel de Aprendizaje Elemental:

... los estudiantes que alcanzan el Nivel de Aprendizaje Elemental demuestran que son capaces de extraer información explícita de acuerdo con datos presentados en tablas, pictogramas (1:1) o gráficos de barras simples, de manera directa o en problemas de un paso.

¿En qué se diferencia el Nivel de Aprendizaje Adecuado del Nivel de Aprendizaje Elemental, en cuanto a las etapas del ciclo de investigación que pueden cubrir?

6. La siguiente actividad simula ser tomada de un texto escolar.

En esta tabla se muestra el número de alumnos inscritos en los diferentes talleres de una escuela.

Taller	Número de alumnos inscritos
Teatro	38
Fútbol	86
Guitarra	45
Pintura	23

¿Cuántos alumnos se inscribieron en el Taller de Guitarra?

a. ¿Cuáles son las etapas del ciclo de investigación que cubre la actividad?
b. Elabore preguntas para que los niños y niñas que se vean enfrentados a la actividad puedan trabajar en todas las etapas del ciclo de investigación.

Capítulo
II

Población, muestra y variables estadísticas

Introducción

En la sociedad en que vivimos, se plantean constantemente preguntas que nos ayudan a conocerla mejor. En ocasiones, es necesario conocer lo que piensa un grupo de personas sobre algún tópico en particular; como, por ejemplo, ¿qué político ha tenido mayor relevancia durante el último año en el mundo?, o responder interrogantes como ¿cuál es la proporción de niños con obesidad hoy en el país?, ¿cuántas familias del primer y segundo quintil de ingresos tienen a sus hijos en escuelas subvencionadas?, entre muchas otras.

Para responder estas preguntas es necesario recolectar datos y extraer de ellos la información. En este proceso, debemos determinar sobre qué o quiénes nos interesa inferir, qué datos se deben recolectar, a qué o a quiénes corresponden estos datos, cuántos datos se necesita recolectar y cómo se realizará dicha recolección, entre otros aspectos; factores que también los niños de Educación Básica deberán enfrentar al querer dar respuesta a las preguntas de investigación presentadas por sus profesores o por ellos mismos.

Para tratar estos aspectos, este capítulo se organiza como sigue: la **Sección 1** muestra, a través de un ejemplo, la necesidad de comprender los conceptos que se estudiarán en las secciones siguientes. La **Sección 2** discute el concepto de población de estudio, mientras que la **Sección 3** se refiere al concepto de muestra. La **Sección 4** discute las diversas formas de recolectar la información requerida para responder a la o las preguntas de investigación. Finalmente, la **Sección 5** se refiere al concepto de variable estadística, y hace una diferenciación entre tipos, lo que será de utilidad para comprender el material desarrollado en los **Capítulos III** y **IV**, que siguen.

1. Motivación

Consideremos, a modo de ejemplo, el estudio estadístico sobre el aprendizaje de las matemáticas "Predictores del desarrollo de conceptos y procedimientos relacionados a las fracciones"[1]. Entre sus principales conclusiones, este estudio señala que existen tanto competencias generales, como relacionadas a aspectos numéricos que resultan importantes para explicar por qué algunos niños tienen grandes problemas con las fracciones. Estas conclusiones deben ser revisadas a la luz de mayor información sobre la manera en que se llevó a cabo el estudio. Un punto fundamental es determinar sobre quién o quiénes se aplican estas conclusiones: ¿niños en qué rango de edades?, ¿en qué niveles escolares?, ¿de todas las realidades socioeconómicas? y ¿sometidos a un tipo particular de enseñanza?, entre otros puntos.

El artículo indica que las conclusiones fueron obtenidas en base a la observación de 357 niños tomados de 9 escuelas de dos distritos de cierto estado en Estados Unidos. El artículo menciona que dichas escuelas incluyen a familias de diversos niveles socioeconómicos. Todos los niños del equivalente a nuestro tercero básico de dichas escuelas fueron invitados a participar, incluyéndose finalmente en el estudio a quienes aceptaron a través de la firma de un consentimiento informado.

Debemos, entonces, preguntarnos a quiénes representan estos niños que participaron en el estudio. Dependiendo de la manera como fueron seleccionadas las escuelas, estos niños pudiesen representar a la región geográfica a la que pertenecen o, más ampliamente, a todos aquellos niños que comparten su nivel de enseñanza y bagaje cultural. Notamos, por ejemplo, que el estudio pudiese no ser aplicable directamente a niños de nuestro país, debido a diferencias de estas índoles. Por otra parte, debemos preguntarnos si el criterio de inclusión, a través de la firma de un consentimiento informado, pudiese o no afectar las conclusiones.

De este modo, si bien los hallazgos reportados por el estudio pueden ser de gran interés, notamos que su rango de validez depende de quiénes fueron los niños incluidos y de la manera en que fueron seleccionados, entre otros aspectos que estudiaremos en este capítulo.

[1] Jordan. N.C. et al. (2013). Developmental predictors of fraction concepts and procedures. *Journal of Experimental Child Psychology*. Vol 116, páginas 45-58.

2. Población

Consideremos el problema de determinar el porcentaje de niños con obesidad en el país. Según discutimos en el capítulo anterior, el planteamiento de este objetivo corresponde al primer paso del ciclo de la investigación.

Para pensar

¿Cuál es exactamente el grupo de niños en el que deseamos detectar la presencia de obesidad?

Debemos especificar, por ejemplo, a qué edades corresponden los niños a los que nos referimos. Debemos indicar si estamos interesados en todos los niños del país o, por ejemplo, solo en los que viven en zonas urbanas, o solo en los que estudian en escuelas municipalizadas, entre otros grupos, lo cual depende del objetivo del estudio. El grupo de interés así definido se denomina *población* en estudio, y debe estar previamente identificado en la pregunta de investigación.

Consideremos, como otro ejemplo, el estudio "Alfabetización en establecimientos chilenos subvencionados"[2] llevado a cabo entre el segundo semestre de 2009 y marzo de 2011. El estudio buscó identificar las condiciones asociadas al aprendizaje de lectura de los niños en establecimientos subvencionados.

Debemos responder, ¿cuál es el grupo de alumnos o escuelas para el cual interesó determinar dichas condiciones? El informe entregado especifica que el estudio estuvo dirigido a alumnos en los niveles de prekinder a segundo básico, en escuelas subvencionadas ubicadas en la Región Metropolitana. Podemos, entonces, inferir que la población de interés corresponde a: *todos los alumnos de escuelas subvencionadas de la Región Metropolitana, que se encuentren cursando dichos niveles escolares.*

En general, podemos decir que la población en estudio es el grupo de sujetos o elementos sobre el cual nos interesa inferir respecto a alguna materia. Una población de interés puede corresponder a los individuos de cierta edad en el país, al estudiar la presencia de una enfermedad; a los individuos de un mismo sexo, al estudiar preferencias de ciertos productos; a un grupo de países con cierta característica, al estudiar calidad de vida, entre otras alternativas.

Es común encontrar que los elementos de interés corresponden a seres vivos, como alumnos, en el estudio sobre alfabetización al que nos referimos previamente, o a plantas, animales o microorganismos, entre otros. Sin embargo, este no es siempre el caso. Por ejemplo, el Ministerio de Salud considera de vital importancia establecer la calidad de los alimentos que entrega a mujeres embarazadas y niños; en particular, se preocupa de la calidad nutritiva de la leche que distribuye. Para esto, selecciona periódicamente algunas cajas de la producción de cada proveedor y estudia su composición alimenticia. En este caso, la población en estudio corresponde a todas las cajas de leche en las bodegas de cada proveedor al momento del estudio.

[2] Tomado de la página web del Plan Nacional del Fomento a la Lectura. http://www.leechilelee.cl/recursos/alfabetizacion-en-establecimientos-chilenos-subvencionados

En la especificación de la población de interés se debe indicar explícitamente, en el caso de que aplique, qué sujetos o elementos se excluyen de esta. A modo de ejemplo, el estudio realizado por el Ministerio de Educación podría, por ejemplo, excluir a todos los alumnos pertenecientes a escuelas de baja matrícula, por considerar que el tipo de instrucción entregada difiere del impartido por la mayoría de las escuelas. En un estudio sobre las preferencias de los vecinos sobre futuros adelantos en una comuna, es posible que la municipalidad desee excluir a propietarios que no residen en ella, entre otros ejemplos. Estas exclusiones deben hacerse explícitas en la descripción de la población en estudio.

Es común encontrar que el objetivo de un estudio consiste en comparar el comportamiento de ciertos subgrupos de sujetos o elementos. A modo de ejemplo, el objetivo de un estudio puede ser comparar el efecto de dos metodologías de estudio diferentes aplicadas sobre dos grupos de niños, o comparar los efectos de un nuevo medicamento sobre pacientes con diferentes niveles de avance de su enfermedad. En este caso, la población de interés corresponde a todos los pacientes que la padecen. Dado que se desea determinar si los efectos de este medicamento dependerán de lo avanzada que se encuentre la enfermedad, el estudio caracterizó las etapas de avance como "inicial", "media", "avanzada" y "muy avanzada". De este modo, será de interés comparar los efectos del medicamento entre pacientes de las diferentes etapas definidas.

En el ejemplo anterior, se han formado grupos de pacientes según la etapa de la enfermedad, donde cada uno de ellos pertenece a uno de los grupos, y donde todos los grupos juntos corresponden al total de las personas que padecen la enfermedad, que hemos definido como la población. Cada uno de estos grupos se denomina una *subpoblación* de la población total, y uno de los objetivos del estudio será comparar los efectos del medicamento entre estas subpoblaciones.

En estudios sobre educación escolar en nuestro país, podríamos comparar los comportamientos de escuelas particulares, subvencionadas y municipales. Cada uno de estos tres tipos de escuela corresponde a una subpoblación de la población de todas las escuelas del país. En general, cuando las subpoblaciones son tales que cada sujeto o elemento puede pertenecer a una y solo una de ellas, estas se denominan *estratos*. Esto significa que las subpoblaciones consideradas no pueden superponerse y que juntas deben corresponder a la población completa.

Ejercicios

1. En cada una de las siguientes situaciones, identifique la población de interés.
   a. Se desea investigar el porcentaje de hogares en el país que ha sido víctima de un asalto durante el último año.
   b. En cierta municipalidad se desea estudiar el ingreso per cápita de los habitantes de la comuna.
   c. Se quiere medir el grado de aprobación de las personas mayores de 21 años a las medidas educacionales planteadas por el gobierno.
2. En cada una de las situaciones anteriores, encuentre dos o más subpoblaciones que pueda ser interesante comparar, de acuerdo al objetivo planteado.

3. Muestra

3.1. ¿Qué es una muestra?

Suponga que la dirección de una escuela desea conocer el menú preferido por los niños para celebrar el Día del Alumno. En este caso, la población en estudio corresponde a todos los alumnos de la escuela. Sin embargo, conocer las preferencias de todos ellos puede resultar complejo. En efecto, supongamos que los niños estudian en dos jornadas diferentes y solo es posible acceder a ellos en una de estas jornadas, o que los niños de distintos niveles utilizan áreas de esparcimiento alejadas, o que, simplemente, no se dispone del tiempo necesario para conocer estas preferencias antes de la celebración. En estos casos, solo podremos acceder a algunos elementos de la población de interés.

Para pensar

En el estudio sobre alfabetización, llevado a cabo por el Ministerio de Educación y mencionado anteriormente, ¿cree que resulta posible acceder a todos los alumnos de la Región Metropolitana? Explique.

Si consideramos el alto número de escuelas y alumnos en el país, y la gran cantidad de tiempo y recursos necesarios para acceder a ellos, un estudio sobre toda la población resultaría imposible.

Sin embargo, podemos acercarnos a lo que ocurre en la población completa observando solo a algunos de sus integrantes. A modo de ejemplo, en el estudio realizado sobre alfabetización, el informe especifica que solo fueron observadas 38 de las escuelas subvencionadas de la Región Metropolitana. Este grupo formado por los niños de las 38 escuelas seleccionadas se denomina *muestra* de la población.

3.2. Representatividad de la muestra

Para pensar

¿Qué condiciones deben darse para que podamos inferir sobre toda la población de interés solo a partir de un subgrupo al que fue posible acceder?

A modo ilustrativo presentamos la **Figura II.1**, que muestra una población de círculos blancos y negros. Nuestro objetivo es determinar el color predominante de círculos en la figura. Si miramos la imagen completa en la **Figura II.1**, es posible decir que predominan los círculos negros. ¿Qué pasaría si solo podemos observar parte de la figura? Situaciones como esta se muestran en las **Figuras II.2** y **II.3**.

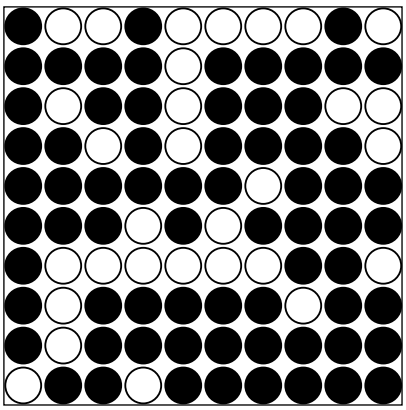


Figura II.1: Población de círculos blancos y negros.

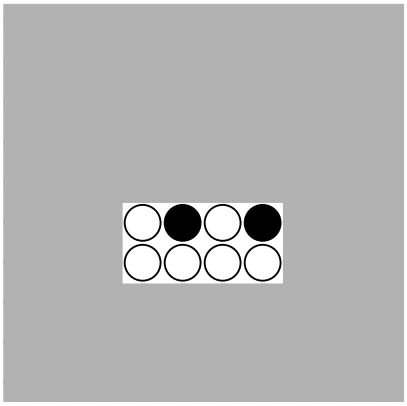


Figura II.2: Muestra de círculos blancos y negros.

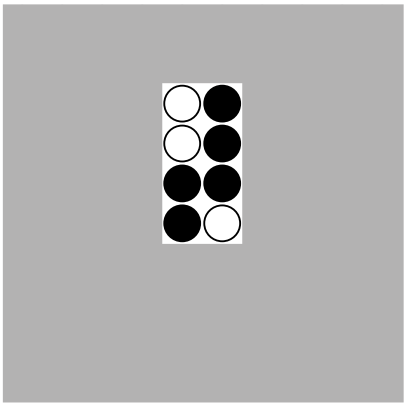


Figura II.3: Muestra de círculos blancos y negros.

En la **Figura II.2**, la sección descubierta muestra el predominio de círculos blancos, lo que podría llevarnos a conclusiones erróneas. La **Figura II.3**, sin embargo, representa de mejor manera el comportamiento de la imagen completa. ¿Qué ocurrirá si miramos otra de estas secciones?

Este ejemplo ilustra la idea de que distintas muestras se comportan de manera diferente o, lo que es lo mismo, que existe *variabilidad entre ellas*. De este modo, las conclusiones obtenidas en base a dos muestras diferentes también pueden diferir. Así nace la necesidad de disponer de uno o más métodos para obtener una muestra, de modo que esta represente de la mejor manera el comportamiento de la población completa. Se dice que una muestra que cumple con esta condición, como en la **Figura II.3**, corresponde a una *muestra representativa* de la población.

En general, una muestra es representativa de una población cuando es similar a esta, respecto a la característica que se desea estudiar. A modo de ejemplo, en el estudio sobre alfabetización es posible que el efecto que se desea determinar, de las prácticas de enseñanza sobre el aprendizaje, dependa fuertemente del nivel socioeconómico que predomine en la escuela. Dado que la población de interés, la Región Metropolitana, posee escuelas de diferentes niveles socioeconómicos, para que la muestra sea representativa, debemos cuidar que todos ellos aparezcan en esta. Por otra parte, un estudio que considere medir la alfabetización de los chilenos no debería solo considerar a los alumnos de escuelas subvencionadas, sino que también debería considerar otros tipos

de escuelas y a la población adulta. Por lo tanto, la muestra elegida por el estudio que seguimos no sería representativa en este caso. Vemos entonces que la representatividad está intrínsecamente ligada a los objetivos de la investigación, por lo que una misma muestra puede ser representativa para un cierto estudio pero no para otro.

Por otra parte, consideremos un estudio realizado por un periódico en línea, respecto de la ley antidiscriminación, donde cada lector debe expresar si cree que la ley debe o no aprobarse. Las respuestas pertenecerán necesariamente al subgrupo de los lectores del periódico, que pueden no representar a la población completa, dado que comparten intereses comunes, y pueden pertenecer mayoritariamente a un solo nivel socioeconómico, entre otras consideraciones. En este caso, los resultados obtenidos no serán representativos de la población y decimos que la muestra es *sesgada*.

Las conclusiones obtenidas a partir de una muestra sesgada no deben ser extrapoladas a la población completa.

Notemos que, hasta acá, hemos hablado de la importancia de contar con una muestra representativa, pero no nos hemos referido a la manera de conseguirla. En las secciones siguientes, se discute este tema en detalle. Sin embargo, adelantamos que una representatividad aproximada se logra seleccionando los elementos de la muestra al azar, entre los elementos de la población, y con un número de elementos seleccionados suficientemente grande.

En resumen

- La *población* corresponde al grupo de sujetos o elementos sobre el cual nos interesa inferir.
- Una *muestra* proveniente de una población corresponde a un subconjunto de esta, sobre el cuál se realizarán las mediciones.
- Una muestra es *representativa* de una población cuando se comporta de manera similar a esta, respecto a la característica de interés en el estudio.
- Una muestra se considera *sesgada* cuando no es representativa de la población, respecto de la característica de interés.

Ejercicios

1. En cada una de las siguientes situaciones, identifique la muestra utilizada:
   a. En cierta municipalidad, se desea estudiar el ingreso per cápita de los habitantes de la comuna. Para ello, varios entrevistadores se sitúan en dos esquinas concurridas de esta y detienen a los transeúntes para preguntarles por sus ingresos.
   b. Se está interesado en medir el grado de aprobación, de las personas mayores de 21 años, a las medidas educacionales planteadas por el gobierno. Para esto, se redacta la afirmación "El aumento de horas de Lenguaje en las escuelas mejorará los puntajes obtenidos en el Sistema de Medición de Calidad de la Educación (SIMCE)", y se le pide a un grupo de

300 personas, contactadas a través de sus teléfonos celulares, que explicite su grado de apoyo a dicha afirmación.

2. En cada una de las situaciones descritas en el ejercicio anterior, discuta la representatividad de la muestra utilizada.

3. Suponga que se desea determinar los hábitos de alimentación típicos de los alumnos en cierta universidad. En cada uno de los siguientes casos, discuta la pertinencia o no de entrevistar solo al grupo mencionado, de modo de obtener una muestra representativa.

   a. Alumnos que comparten un departamento.
   b. Alumnos de la especialidad de Salud Pública.
   c. Alumnos que participan en selecciones deportivas.
   d. Alumnos inscritos en una clase obligatoria de inglés.

3.3. Tipos de muestreo

Como mencionamos anteriormente, para determinar si una muestra es representativa de una población, debemos estudiar cómo fue elegida. Supongamos que se desea conocer cierta característica de los niños de un curso. En este caso, la población de interés corresponde al curso completo, pero realizaremos el estudio encuestando solo a 10 niños. Para ello, considere la siguiente propuesta: se ingresan en una urna (bolsa, caja, sobre, etc.) tantas papeletas cerradas como niños tenga el curso, y se escriben previamente los nombres de los niños en las papeletas. Luego, pedimos a uno o varios niños que elijan una papeleta desde la urna, sin haberlas visto previamente, hasta alcanzar 10 papeletas en total. Los nombres de los niños en las papeletas seleccionadas corresponden a los nombres de los niños seleccionados en la muestra.

La forma de selección propuesta entregará una muestra representativa de los niños del curso, pues todos ellos tenían las mismas oportunidades de ser elegidos. Cuando esto ocurre, decimos que la selección fue hecha mediante un muestreo *aleatorio simple*.

Para pensar

¿Cree usted que la muestra de niños elegida a través de papeletas que se encuentran dentro de una urna sigue siendo representativa si alguna de las papeletas fuera más grande? ¿Y si la urna es transparente y las papeletas son de diferentes colores?

En el primer caso, las papeletas de mayor tamaño tienen más posibilidades de ser elegidas. En el segundo, la muestra puede estar sesgada hacia los colores favoritos de los niños que eligen las papeletas. En este caso, las muestras no serían representativas del curso, pues no todos los niños tendrían las mismas oportunidades de ser elegidos.

Otra manera de definir un muestreo aleatorio simple es considerar todas las muestras posibles de un tamaño dado, que se podrían generar a partir de una población. El muestreo aleatorio simple elige una muestra, entre todas estas muestras, asignando a cada una las mismas posibilidades de ser elegidas.

A modo de ejemplo, consideremos una población pequeña de 5 niños, Antonia, Pedro, Ernesto, Andrés y Carla, y supongamos que deseamos tomar una muestra aleatoria simple de 2 niños. La **Tabla II.1** contiene las 10 posibles muestras de 2 niños:

Muestra	Integrantes de la muestra
1	Antonia, Pedro
2	Antonia, Ernesto
3	Antonia, Andrés
4	Antonia, Carla
5	Pedro, Ernesto
6	Pedro, Andrés
7	Pedro, Carla
8	Ernesto, Andrés
9	Ernesto, Carla
10	Andrés, Carla

Tabla II.1: Lista de las 10 posibles muestras de dos integrantes a partir de la población de 5 niños.

El procedimiento consistirá en escribir 10 papeletas con los números del 1 al 10, y se elegirá solo una de estas papeletas de manera aleatoria. Supongamos que elegimos la papeleta con el número 3. Entonces, de la **Tabla II.1** leemos que la muestra elegida estará compuesta por Antonia y Andrés.

El mismo procedimiento aplicado a un problema similar se ilustra en las **Figuras II.4**, **II.5** y **II.6**. En ellas, se ilustra la elección de una muestra aleatoria simple de 2 letras, a partir de las letras A, B, C, D y E.

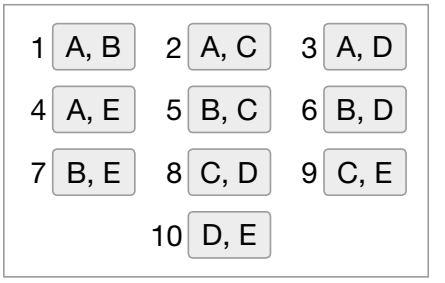


Figura II.4: Lista de todas las muestras posibles de tamaño 2 enumeradas del 1 al 10, a partir de las letras A, B, C, D y E.

Figura II.5: Se ingresan en la urna papeletas con los números del 1 al 10, y se elige de manera aleatoria una de ellas. Se ha elegido la muestra número 3.

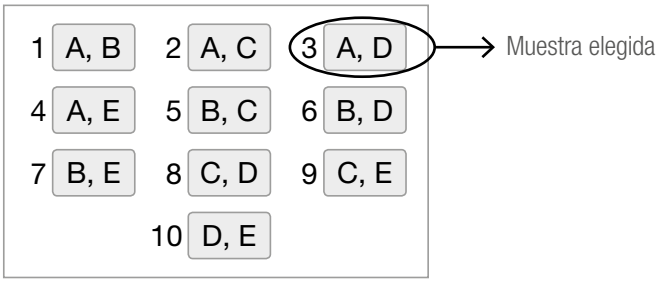


Figura II.6: La muestra elegida corresponde a la muestra formada por las letras A y D.

Si bien esta metodología ilustra una de las definiciones del muestreo aleatorio simple, la manera más simple de obtenerla corresponde a la primera metodología descrita. Es decir, para elegir una muestra aleatoria simple de 2 niños, entre Antonia, Pedro, Ernesto, Andrés y Carla, ingresamos en una urna 5 papeletas, una con cada uno de los nombres de los niños, y sacamos dos papeletas para elegir a los 2 niños que compondrán la muestra.

Los dos mecanismos descritos: elegir tantas papeletas de una urna, como niños se quiera en la muestra, o elegir la muestra completa, como se muestra en las **Figuras II.4**, **II.5** y **II.6**, son procedimientos equivalentes para realizar un muestreo aleatorio simple.

En la práctica, cuando las poblaciones y/o las muestras son de mayor tamaño y se hace físicamente imposible implementar cualquiera de los procedimientos descritos, la selección puede realizarse utilizando un programa computacional. Sin embargo, al enseñar este tópico en el aula, es importante comenzar con la implementación física del método, para facilitar la comprensión de los niños.

Una de las características del muestreo aleatorio simple es que el experimentador controla las oportunidades de salir de cada elemento o individuo en la población. Cuando un muestreo presenta estas características, decimos que el muestreo es *probabilístico*. Un muestreo probabilístico es adecuado cuando el experimentador controla estas oportunidades de manera que garanticen la representatividad de la muestra.

En este contexto, ya vimos que en el muestreo aleatorio simple las oportunidades de cada individuo o elemento de la población de pertenecer a la muestra son las mismas. Sin embargo, existen situaciones donde no necesariamente se quiere que esto ocurra. Así se generan otros

tipos de muestreo, también probabilísticos, como el muestreo *estratificado* o el muestreo *por conglomerados*. El primero de ellos se utiliza cuando existen otras variables, diferentes a las variables en estudio, que pudiesen afectar los resultados del estudio. El segundo, cuando existen dificultades de acceso a ciertos subgrupos de la población completa. Ninguno de estos tipos de muestreo sacrifica la representatividad de la muestra.

Un ejemplo clásico de muestreo *no probabilístico* corresponde a uno donde los sujetos deciden pertenecer a la muestra por iniciativa propia. Esto se denomina *autoselección*, y se da, por ejemplo, en las encuestas realizadas por programas de televisión, revistas o diarios, donde los individuos que las responden no han sido previamente seleccionados, sino que son ellos los que han decidido participar en la encuesta. Se ha estudiado que las personas tienen mayor motivación para emitir su opinión cuando esta es desfavorable con respecto al tópico de interés. De este modo, las conclusiones obtenidas a través de un muestreo de este tipo pueden resultar más pesimistas de lo que son, en realidad, en la población. Otro ejemplo de muestreo no probabilístico corresponde a la entrevista de transeúntes en cierto lugar de la ciudad. Es altamente probable que este tipo de muestras resulte no ser representativa de la población. Este muestreo se denomina muestreo *por conveniencia*.

En general, una muestra pequeña, pero elegida según un muestreo probabilístico, será de mejor calidad que una muestra de gran tamaño elegida a través de un muestreo no probabilístico, como el muestreo por conveniencia.

En resumen

- El tipo de muestreo utilizado determina la representatividad de la muestra obtenida.
- Un *muestreo probabilístico* controla las posibilidades de los individuos de pertenecer a la muestra. Si estas son controladas de manera correcta, la muestra obtenida será representativa.
- El *muestreo aleatorio simple* corresponde a un muestreo probabilístico donde todos los individuos tienen la misma oportunidad de ser elegidos.

Ejercicios

1. En las siguientes situaciones, indique si el tipo de muestreo utilizado es probabilístico. En caso de serlo, identifique si, en alguna de sus partes, se realizó un muestreo aleatorio simple.
   a. Se realiza un test del sabor de unos nachos y una salsa a personas que ingresan a un supermercado entre las once y las doce de la mañana.
   b. Un computador selecciona aleatoriamente los números de matrícula de **100** alumnos de una escuela para preguntarles por sus pasatiempos.
   c. Con el objetivo de conocer el grado de satisfacción laboral de los profesores de escuelas subvencionadas, se escoge de manera aleatoria una muestra de **15** escuelas y se entrevista a todos los profesores en cada una de ellas.

2. Considere nuevamente el problema de elegir una muestra de 2 niños a partir de una población compuesta por 5 niños, Antonia, Pedro, Ernesto, Andrés y Carla. En este ejercicio intentaremos verificar empíricamente que las dos formas propuestas para realizar un muestreo aleatorio simple son equivalentes.

   - La **Tabla II.1**, que se presenta al comienzo de esta subsección, lista las 10 posibles muestras de 2 niños a partir de la población de 5. Recorte 10 papeletas, y en cada una de ellas anote una de las muestras posibles (cada papeleta incluirá los nombres de 2 niños). Ingrese estas papeletas en una bolsa o caja, donde no se pueda visualizar el interior.
   - Sin mirar el interior del recipiente, elija una papeleta y registre la muestra que contiene.
   - Devuelva la papeleta a la urna, y repita el procedimiento 20 veces. Note que, como cada papeleta contiene los nombres de 2 niños, en total habrá extraído 40 nombres (con repeticiones de ellos).
   - Una vez que tenga las 20 muestras, complete una tabla como la siguiente:

Nombre	Número de veces que apareció en una muestra
Antonia	
Pedro	
Ernesto	
Andrés	
Carla	
Total	40

   En base a la tabla que ha llenado, ¿cree usted que todos los niños tenían las mismas posibilidades de pertenecer a la muestra?

3.4. Algunos errores relacionados al concepto de muestra

Algunas de las concepciones erróneas que suelen aparecer están relacionadas con la calidad de una muestra. A continuación discutimos algunos de estos errores.

- *Creer que una muestra de mayor tamaño siempre es mejor*. Esto es cierto solo cuando se está utilizando un método de muestreo adecuado. En el caso de realizar una encuesta a televidentes de un programa, donde cada uno de ellos responde de manera voluntaria, la calidad de una muestra de gran tamaño será peor que la calidad de una muestra de menor tamaño, pero elegida según un tipo de muestreo probabilístico.
- *Creer que la precisión de las inferencias depende del porcentaje o fracción de la población al que corresponde la muestra*. Cuando el tamaño de la población es grande, la precisión de las inferencias está dada exclusivamente por el tamaño de la muestra en sí mismo, y no por este en relación al tamaño de la población completa.

- *Creer que un censo es mejor que una muestra aleatoria representativa*. Si bien un censo permite acceder a toda la población, una muestra aleatoria tomada de manera adecuada, asegurando su representatividad, puede entregarnos información muy precisa, a un costo muchísimo menor.
- *Creer que se puede obtener una muestra representativa sin utilizar un método aleatorio*. Existe la creencia de que una buena metodología para elegir una muestra representativa corresponde a elegir de manera arbitraria a personas que "creemos" que son representativas de la población en términos de alguna característica. Este procedimiento no garantiza la representatividad de la muestra, dado que incluye un componente subjetivo.

4. Recolección de datos e información

Luego de establecer la pregunta y la población de estudio, la siguiente etapa es recolectar los datos y la información pertinente sobre los atributos o cualidades de los participantes: casos, comunidades u objetos involucrados en la investigación.

La recolección de datos e información implica elaborar un plan detallado de procedimientos que permitan reunir los datos necesarios para dar respuesta a la pregunta. Este plan incluye responder las siguientes preguntas:

1. **¿Cuáles son las fuentes desde donde se obtendrán los datos?**

Se debe determinar si los datos serán proporcionados por personas, a partir de la observación de experimentos o simulaciones, o a partir de información almacenada en documentos, archivos o bases de datos, entre otras fuentes. La selección de la o las fuentes dependerá de las preguntas de investigación, los fenómenos que se quieren estudiar y el conocimiento de los investigadores. A modo de ejemplo, dado el tipo de preguntas de investigación que los niños plantean, alumnos de primero básico debiesen centrar tales fuentes en individuos pertenecientes a contextos cercanos, como la familia o el mismo curso, o bien en experimentos simples, como la observación del crecimiento de una planta o el lanzamiento de dados. Al avanzar en la Educación Básica, los alumnos propondrán preguntas más complejas, con lo que las fuentes podrían ampliarse a todos los alumnos de la escuela, o a individuos del barrio o de la comunidad. De esta misma forma, los experimentos se complejizan en conocimiento y tipos de tareas para identificar los datos.

2. **¿Dónde se localizan tales fuentes?**

Una vez identificadas las fuentes, se debe determinar dónde y cómo se accederá a ellas. En esta etapa, se debe reflexionar sobre las oportunidades que tiene el investigador para esto. El análisis de los alcances y limitaciones de estas oportunidades permite al investigador conocer la factibilidad del estudio, o algunas restricciones que este posea. A modo de ejemplo, si alumnos de cuarto básico desean estudiar los hábitos de estudio de los alumnos de toda la escuela, debieran reflexionar sobre condiciones de acceso, como el tiempo que ellos posean para entrevistar a todos los alumnos o el horario de disponibilidad de estos, entre otros aspectos.

3. **¿A través de qué medio o método se van a recolectar los datos?**

Esta fase del plan implica elegir uno o varios medios de recolección de datos y definir los procedimientos que se utilizarán en ellos. Estos métodos deben ser confiables, válidos y objetivos. El

método más utilizado para la recolección de datos e información es el cuestionario, que consiste en un conjunto de preguntas que sirven para obtener la información necesaria para el estudio. Un ejemplo de cuestionario se muestra en la Figura II.7.

Estudio de satisfacción

Propósito: la siguiente encuesta tiene como propósito recabar información respecto de la satisfacción que los alumnos de la escuela tienen sobre la calidad de los alimentos que el quiosco ofrece como colación.

Preguntas

A continuación, se presenta una lista de preguntas sobre la satisfacción que tienes respecto de la calidad de los alimentos que el quiosco de la escuela vende para las colaciones. Por favor, marca con una **X** la respuesta que mejor te represente.

M: Muy satisfecho B: Bastante satisfecho P: Poco satisfecho I: Insatisfecho

Pregunta	M	B	P	I
¿Estás satisfecho con los platos de fondo que el quiosco ofrece?				
¿Estás satisfecho con los postres que el quiosco ofrece?				
¿Estás satisfecho con las frutas que el quiosco ofrece?				
¿Estás satisfecho con los jugos que el quiosco ofrece?				

Figura II.7: Ejemplo de cuestionario utilizado para obtener información relevante para cierto estudio.

En un cuestionario pueden haber distintos tipos de preguntas: preguntas abiertas, donde el sujeto al cual se le está aplicando el cuestionario puede expresar libremente lo que piensa u opina respecto al tema de la pregunta; o preguntas cerradas, las cuales contienen categorías u opciones de respuesta.

Un ejemplo de pregunta abierta es:

¿Qué te motiva a estudiar Pedagogía en Educación Básica?

Un ejemplo de pregunta cerrada es:

¿Te gusta el color verde?

( ) Sí ( ) No

El uso de uno u otro tipo de pregunta depende del plan para recolectar los datos y la información, de la forma en la que el cuestionario se aplique y de la pregunta de investigación. Por ejemplo, consideremos la siguiente pregunta de investigación: ¿qué tan saludables son las personas que integran la comunidad educativa? Para saber esto, se deberían caracterizar los elementos propios de una persona saludable, para luego consultarlos a los sujetos de la población de estudio. Una de las preguntas podría ser: ¿usted cree que es una persona saludable? ¿por qué sí o por qué no?, donde se quiere saber la percepción de los sujetos sobre qué es ser saludable para ellos. Pero también podría haber preguntas cerradas, como, por ejemplo: aproximadamente, ¿cuántas horas a la semana dedica usted a la actividad física?, entregando las alternativas: Ninguna, Entre 1 y 5, y Más de 5.

La decisión sobre el tipo de pregunta a utilizar, abierta o cerrada, resulta relevante a la hora de extraer la información. A modo de ejemplo, al realizar una encuesta de opinión, la práctica ha demostrado que los resultados obtenidos dependen fuertemente de la manera en que haya sido planteada la pregunta. Un caso descrito en la literatura [3] muestra el efecto de utilizar dos preguntas diferentes para obtener información sobre los problemas de interés nacional de mayor importancia para los ciudadanos. Una de ellas corresponde a una pregunta abierta: "¿cuál piensa usted que es el problema más importante que enfrenta el país hoy en día?". La segunda redacción definió la pregunta como cerrada: "¿cuál de los siguientes cree usted que es el mayor problema que enfrenta nuestro país hoy en día?", listando las alternativas: Energía, Educación escolar, Aborto, Contaminación, Otros y No sabe.

Las respuestas se describen en la **Tabla II.2**. En ella se ha listado únicamente las alternativas de respuesta utilizadas en la pregunta cerrada. En la tabla se observan diferentes resultados frente a la misma inquietud, dependiendo del tipo de pregunta utilizado. La información obtenida a partir de ambos estudios conduce a diferentes conclusiones.

Problema	Pregunta abierta	Pregunta cerrada
Energía	0,0%	5,6%
Educación escolar	1,2%	32,0%
Aborto	0,0%	8,4%
Contaminación	1,1%	14,0%
Otros	93,0%	39,4%
No sabe	4,7%	0,6%

Tabla II.2: Respuestas obtenidas a través de dos preguntas diferentes dirigidas a identificar los problemas más importantes que enfrenta el país.

A partir de la segunda columna de la **Tabla II.2**, vemos que dejar la pregunta abierta nos llevó a encontrar que, para un gran porcentaje de la población (93%), los problemas reportados no correspondían a los problemas listados en la pregunta cerrada. Esto muestra una falencia de utilizar una pregunta cerrada: es posible que esta ignore una de las alternativas más comunes, lo que esconderá posible información relevante.

Por otra parte, existen dos formas base para la aplicación de un cuestionario o encuesta:

- *Autoadministrado*: el cuestionario se entrega a los sujetos participantes del estudio y ellos lo contestan directamente, sin ningún medio o filtro.
- *Por entrevista*: el cuestionario se aplica a través de un filtro: una persona calificada hace las preguntas a los sujetos y anota las respuestas.

[3] Moore, D. y Notz, W.D (2012), *Statistics: Concepts and Controversies*. McGraw Hill. Décima edición.

En los primeros años de Educación Básica, los alumnos acostumbran aplicar los cuestionarios mediante entrevista, registrando las respuestas en tablas de conteo o registro. A medida que avanzan en Educación Básica, podrían utilizar cuestionarios autoadministrados, ya que, teniendo mayores habilidades en el área de lenguaje y comunicación, serán capaces tanto de construirlos, como de posteriormente extraer la información que requieren a partir de estos.

Otra forma de recolección de información es la observación. La observación es el registro visual de lo que ocurre en una situación, consignando y clasificando los datos con algún esquema previsto, como por ejemplo, una tabla. Estas situaciones pueden ser experimentos o simulaciones. Un ejemplo de esta forma de recolección de datos puede ser la cuantificación de autos que transitan en una esquina a través de la observación en el lugar.

Ejercicio

Considere un estudio para conocer la opinión de los habitantes del país sobre el financiamiento de campañas electorales. Se presentan dos maneras de redacción de la pregunta:

- ¿Se deben aprobar leyes en el país que eliminen toda posibilidad de que grupos de interés aporten grandes sumas de dinero a las campañas electorales?
- ¿Se deben aprobar leyes para prohibir que grupos de interés contribuyan a las campañas electorales, o estos grupos tienen el derecho de contribuir al candidato que apoyan?

Compare las dos redacciones y discuta los efectos que pueden tener sobre las respuestas de los entrevistados.

5. Parámetro y estadístico

Cuando diseñamos un estudio debemos tener presente de qué manera vamos resumir o caracterizar el comportamiento de los atributos o características de interés, tanto en la muestra como en la población.

A modo de ejemplo, supongamos que nos interesa estudiar la altura de los niños de los tres terceros básicos de una escuela. Podemos representar aspectos de esta variable a través de ciertas cantidades relevantes, como, por ejemplo, la altura promedio de todos los niños. Esta cantidad corresponde a una característica de la población completa, lo que se denomina un *parámetro poblacional*. Dado que, en general, es imposible conocer la población completa, no es posible tampoco conocer el valor del parámetro. Esto caracteriza a un parámetro poblacional como algo no observable.

Sin embargo, podemos utilizar una muestra obtenida para tener una idea del valor de un parámetro poblacional. De este modo, el promedio de las alturas de los niños en una muestra puede servirnos para inferir sobre la altura promedio de todos los niños de los terceros básicos. El promedio de las alturas en la muestra se denomina un *estadístico muestral* y, a diferencia de un parámetro poblacional, es observable. La **Figura II.8** ilustra esta situación. En ella se observa una población de anillos negros y blancos, donde el objetivo es conocer la proporción de anillos blancos. Esta proporción corresponde a un parámetro de la población de anillos de colores y, dado el tamaño de la población y la imposibilidad de observarla completamente, este tampoco es observable.

Sin embargo, se toma una muestra de 5 anillos, como se ve en la imagen de la derecha, en la **Figura II.8**, calculándose en ella la proporción de anillos blancos. Esta proporción corresponde a un estadístico muestral, cuyo valor es observable y, en este caso, corresponde a $\frac{2}{5}$ (2 anillos blancos de 5 anillos en total).

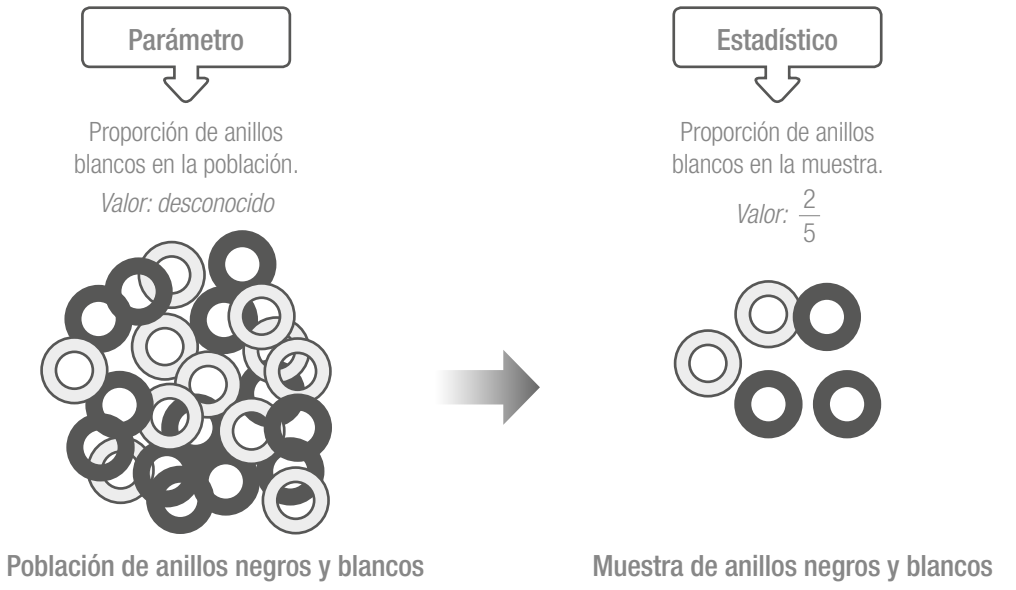


Figura II.8: Conceptos de parámetro poblacional y estadístico muestral.

Un error es confundir los valores obtenidos en la muestra, o estadísticos, con los valores en la población, o parámetros. Una idea para recordar es que, como veremos más adelante, utilizaremos los valores de los estadísticos de la muestra para estimar los correspondientes parámetros de la población.

Así, en el ejemplo representado en la **Figura II.8**, si la muestra ha sido tomada a través de un muestreo probabilístico adecuado, podríamos inferir que el valor del parámetro poblacional, es decir, de la proporción de anillos blancos en la población de anillos de colores, es *cercana* a $\frac{2}{5}$.

Debemos tener cuidado al hacer afirmaciones como la anterior, ya que, aunque resultan intuitivas, presentan incerteza. Esta incerteza depende tanto de cómo fue tomada la muestra, a lo que nos referimos en secciones anteriores, como del tamaño de esta, como veremos en ejemplos que siguen.

Es posible verificar que, al tomar diferentes muestras, un mismo estadístico muestral tomará diferentes valores, pero estos seguirán un patrón. A modo de ejemplo, consideremos una población de 5 niños y supongamos que interesa conocer el promedio del número de hermanos que tiene cada uno. La **Tabla II.3** muestra el número de hermanos de cada uno de ellos. El número promedio de hermanos de todos los niños corresponde a:

$$\frac{4+1+0+5+0}{5} = 2$$

Niños	Número de hermanos
Sebastián	4
Isabel	1
Catalina	0
Josefa	5
Mónica	0

Tabla II.3: Número de hermanos de todos los niños de la población.

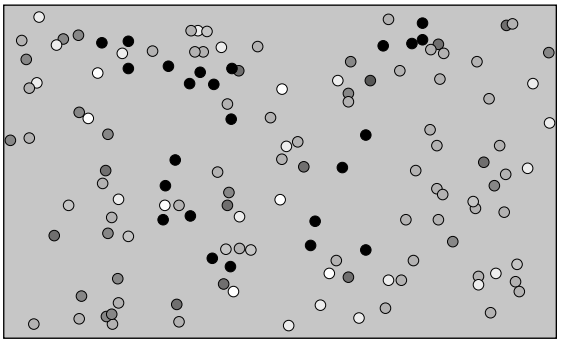
La **Tabla II.4** muestra todas las posibles muestras de tamaño **2** provenientes de esta población y el promedio de hermanos observado en cada una de estas muestras. Así, se tiene que mientras el parámetro poblacional es fijo (e igual a **2** hermanos), los estadísticos muestrales, o promedios en cada muestra, cambian de valor de una muestra a otra.

Niños en la muestra	Suma del número de hermanos	Promedio
Sebastián e Isabel	4 + 1	2,5
Sebastián y Catalina	4 + 0	2,0
Sebastián y Josefa	4 + 5	4,5
Sebastián y Mónica	4 + 0	2,0
Isabel y Catalina	1 + 0	0,5
Isabel y Josefa	1 + 5	3,0
Isabel y Mónica	1 + 0	0,5
Catalina y Josefa	0 + 5	2,5
Catalina y Mónica	0 + 0	0
Josefa y Mónica	5 + 0	2,5

Tabla II.4: Todas las muestras posibles de 2 integrantes, a partir de la población de 5 niños y sus promedios de número de hermanos. El valor del parámetro poblacional, o promedio de los 5 niños, es 2.

Si bien esperaríamos que los promedios del número de hermanos de los niños en cada muestra fueran cercanos al promedio de los **5** niños (**2** hermanos), observamos que no siempre es cierto. Esto se debe a que el tamaño de la muestra, **2** niños, es bastante pequeño, lo que se traduce en incerteza sobre los resultados.

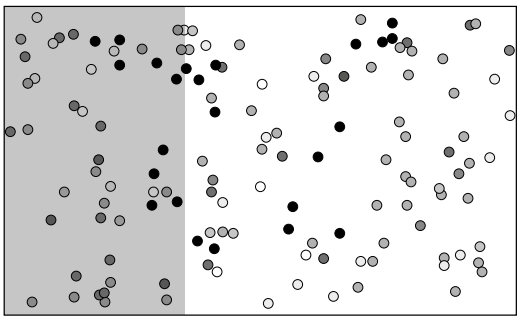
En otro ejemplo, la **Figura II.9** muestra los **136** niños de los **3** terceros básicos de una escuela. Cada círculo corresponde a un niño de estos cursos. El color de cada círculo representa su altura, correspondiendo colores más intensos a niños de mayor altura. El parámetro poblacional corresponde al promedio de la altura de todos los niños. A través de un programa computacional, es posible saber el promedio de altura de todos los niños de los terceros básicos, que es **135,49 cm**.



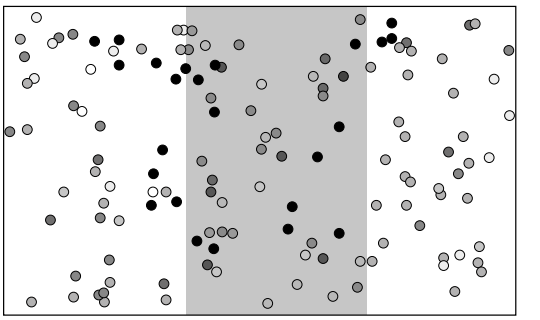
Altura promedio: 135,49 cm.

Figura II.9: Población de niños de los terceros básicos de una escuela. La altura promedio de todos los niños es 135,49 cm, lo que corresponde a un parámetro poblacional.

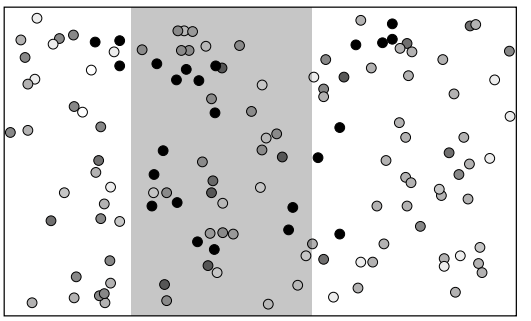
Por otra parte, en la **Figura II.10** se muestran 4 muestras aleatorias probabilísticas diferentes de 47 niños cada una. En cada una de ellas, se ha calculado la altura promedio de los niños que le pertenecen, las cuales se indican al pie de cada figura. Observamos que las alturas promedio de los niños en cada una de las 4 muestras obtenidas son diferentes, pero que, sin embargo, se mueven entre **134,91 cm** y **135,97 cm**. Estos valores se acercan bastante al valor del parámetro poblacional o promedio de alturas de todos los niños, de **135,49 cm**. A diferencia de lo que observamos en el ejemplo del número de hermanos, donde los promedios de las muestras presentaban gran variabilidad, en este ejemplo la variabilidad de las alturas promedio en las muestras es bastante menor. Esto se debe a que, en este último caso, las muestras consideradas son de mayor tamaño.



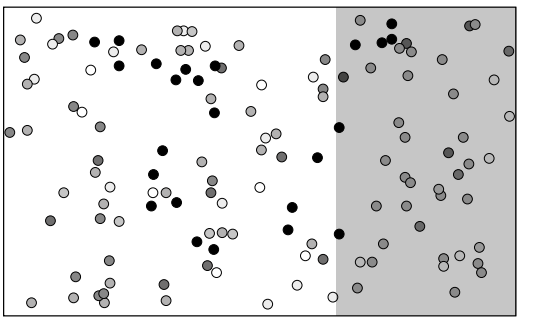
Altura promedio: 135,97 cm.



Altura promedio: 135,27 cm.



Altura promedio: 134,91 cm.



Altura promedio: 135,70 cm.

Figura II.10: 4 muestras de niños de los terceros básicos. Las secciones grises encierran los niños de cada muestra, y se indican sus alturas promedio (estadísticos muestrales). El promedio de todos los niños (parámetro poblacional) es 135,49 cm.

Este ejemplo ilustra que es posible acercarse a los valores de los parámetros poblacionales a través de los valores que toman algunos estadísticos muestrales. Para que este procedimiento sea adecuado, se debe asegurar que la muestra tomada sea probabilística y que el número de observaciones sea suficientemente grande. Esto permitirá hacer inferencias sobre la población en estudio, y es lo que permitirá dar respuesta a la o las preguntas de investigación.

En los **Capítulos III** y **IV**, donde nos referimos a diferentes formas de representar y resumir datos, utilizaremos el concepto de *conjunto de datos*, en lugar de muestra, debido a que, en algunas ocasiones, como, por ejemplo, al trabajar con niños en los primeros años de escolaridad, los datos recolectados pertenecen a la población de interés completa, como su curso, sus familias o amigos. En estos casos, las representaciones y resúmenes entregados corresponden a

parámetros poblacionales. Sin embargo, esta situación rara vez se da en la práctica, y las representaciones o resúmenes de datos obtenidos corresponden a estadísticos muestrales, siendo el objetivo principal utilizarlos para realizar inferencias sobre la población.

En resumen

- Los *parámetros* corresponden a aspectos de la población y no son observables.
- Los *estadísticos* corresponden a aspectos de la muestra y son observables.
- Podemos aprender sobre el valor de un parámetro poblacional utilizando estadísticos muestrales.

Podemos integrar todos los conceptos que hemos estudiado notando el lugar en que estos aparecen en el ciclo de la investigación. La **Figura II.11** corresponde a la representación del ciclo de investigación que dimos en el **Capítulo I**, incluyendo ahora leyendas que indican el lugar donde hace su aparición cada uno de los conceptos aprendidos en este capítulo. En la etapa del *Planteamiento del problema*, la pregunta de investigación debe indicar claramente cuál es la población en estudio. La etapa de *Planificación* contempla determinar la forma de recolectar la información y cómo se selecciona los individuos o elementos que constituirán la muestra. En la etapa de *Recolección de datos*, se pone en práctica cada uno de los puntos convenidos en la *Planificación*. La etapa de *Análisis* incluye la obtención de estadísticos muestrales, mientras que, finalmente, en la etapa de *Conclusiones* se utilizan los resultados del análisis para realizar inferencias sobre los parámetros poblacionales de interés.

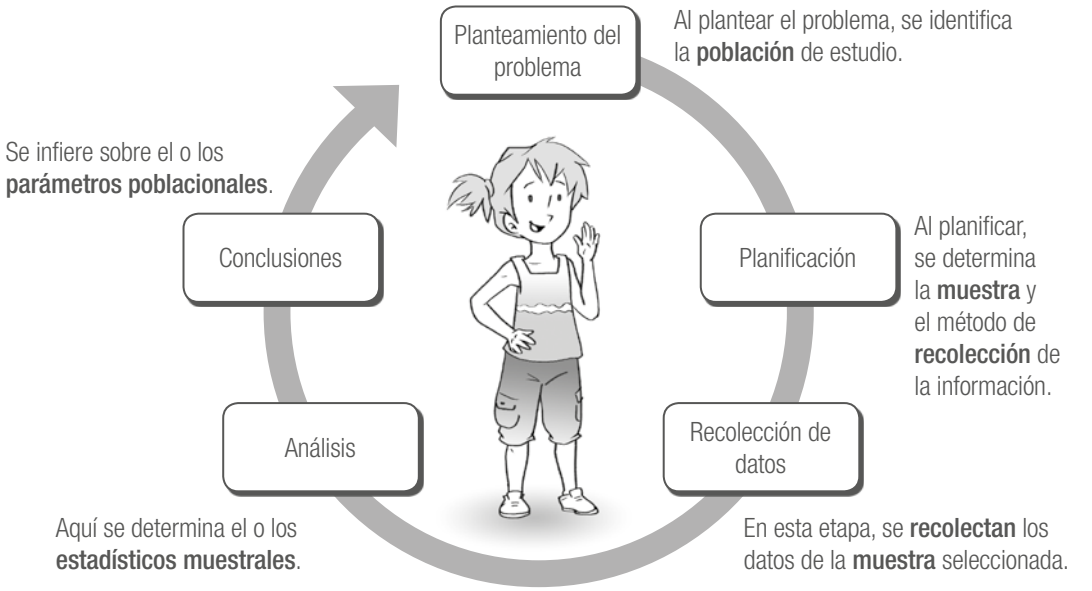


Figura II.11: Los conceptos de población, muestra, estadísticos muestrales y parámetros poblacionales, en las etapas del ciclo de la investigación.

Ejercicio

En los siguientes apartados, indique si la cantidad resaltada en cursiva corresponde a un parámetro poblacional o a un estadístico muestral. Justifique.

a. El Ministerio del Trabajo anunció que el mes anterior se entrevistó a todos los miembros de la fuerza laboral en una muestra de **60.000** hogares: un ***6,5%*** de las personas entrevistadas estaban desempleadas.

b. Un sistema de ventas telefónicas utiliza un dispositivo que marca números telefónicos residenciales de manera aleatoria. Entre los primeros **100** números marcados, ***43*** no correspondían a números encontrados en la guía telefónica. Este resultado no es sorprendente, ya que un ***52%*** de todos los teléfonos residenciales del país no se encuentran en dicha guía.

c. Un lote de rodamientos producidos tiene un diámetro medio de ***2,53*** **cm**, lo cual se encuentra dentro de los límites de las especificaciones. Un inspector elige **100** de estos rodamientos en el lote y obtiene un diámetro medio de ***2,61*** **cm**.

6. Variables estadísticas

Consideremos, a modo de ejemplo, un estudio que desea inferir acerca de la obesidad de los niños en el país. Una vez escogidos los niños, quisiéramos conocer su peso y talla. Posiblemente, otra característica de interés del estudio, por encontrarse relacionada con la obesidad, puede ser el tiempo semanal que cada niño dedica a la actividad física. Interesará también conocer el sexo del niño y sus características sociales y demográficas: el número de hermanos, la comuna de residencia y el nivel socioeconómico, entre otra información.

Cada una de las características o atributos que hemos mencionado, como peso, talla, tiempo semanal que dedica a la actividad física, sexo, número de hermanos, comuna o nivel socioeconómico, se denomina *variable estadística*.

En el estudio "Alfabetización en establecimientos chilenos subvencionados", donde se desea estudiar el efecto de ciertas políticas educacionales sobre el rendimiento de los niños en escuelas subvencionadas, la principal variable de interés corresponde al rendimiento de cada alumno en una evaluación de su capacidad lectora. Otras variables posiblemente relevantes corresponden al nivel que cursa el alumno, la escuela en que está, el tamaño de su curso o su edad, entre otras.

Las variables que hemos mencionado son de diferente índole. A modo de ejemplo, las variables sexo, escuela, comuna y nivel socioeconómico toman valores que corresponden a categorías, no asociadas a un valor numérico. Las categorías de la variable sexo corresponden a "hombre" y "mujer", las categorías de la variable escuela corresponden a todos los nombres de las escuelas en el país, y así con las variables comuna y nivel socioeconómico. Este tipo de variable, cuyos valores posibles no son numéricos, se denominan *variables cualitativas*, por corresponder a cualidades de los individuos o elementos en estudio.

Sin embargo, también existen diferencias entre las variables cualitativas. Notemos que las categorías de la variable sexo no tienen orden: es irrelevante ordenarlas como hombre y mujer o viceversa. Lo mismo ocurre para las categorías de la variable escuela y comuna: no existe un orden natural para ellas, sino que solo nos interesa el nombre de la escuela o la comuna, según corresponda. Este tipo de variables se denomina *cualitativa nominal*.

Por otra parte, las categorías de la variable nivel socioeconómico pueden ser ordenadas, a modo de ejemplo, como ABC1, C2, C3, D, E. Estas categorías son ordenables, ya sea en el orden en que han sido presentadas, de mayor a menor nivel socioeconómico, o en el inverso. Las variables con esta propiedad se denominan variables *cualitativas ordinales*, ya que se pueden ordenar sus categorías.

Consideremos ahora variables como peso, talla, tiempo semanal que dedica a la actividad física y número de hermanos. Cada una de ellas toma solo valores numéricos. Así, estas variables se denominan *cuantitativas*, por corresponder a cantidades.

Nuevamente notemos diferencias entre estas. Variables como, peso, talla y tiempo dedicado a la actividad física se caracterizan por que, entre dos valores cualesquiera, tiene sentido hablar de un valor intermedio. A modo de ejemplo, aun cuando midamos el tiempo con un cronómetro cuya precisión sean los segundos y no sea posible obtener fracciones de ellos, podemos pensar que el tiempo real exacto sí puede tomar valores intermedios y que el cronómetro es únicamente el instrumento que se ha utilizado para medirlo de manera aproximada. En situaciones donde esto ocurre,

se dice que la variable es *cuantitativa continua*. La condición descrita no se cumple, por ejemplo, para variables como el número de hermanos. En efecto, si tomamos dos valores cualesquiera de ella, como 3 y 4 hermanos, no existe un valor intermedio válido para la variable, dado que no se puede tener, por ejemplo, 3,2 hermanos. Cuando esto ocurre, se dice que la variable es *cuantitativa discreta*.

La **Figura II.12** ilustra la clasificación de variables que hemos hecho:

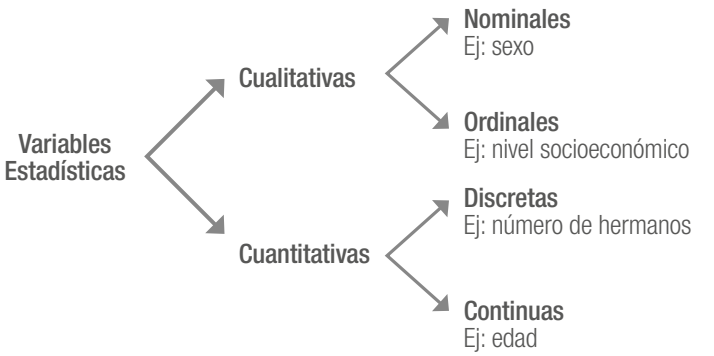


Figura II.12: Tipos de variables estadísticas.

Para pensar

En una muestra mediremos el peso en kilos. La variable peso, ¿se considera discreta o continua?

Si pensamos que el peso será medido en kilos, podemos decir que solo tomará valores enteros, por lo que la variable peso sería considerada discreta. Incluso, si la consideramos con mayor precisión, midiéndola en gramos, aun nos seguiría pareciendo discreta por no poder observar ningún peso como, por ejemplo, **21,3 gramos**, pues solo observamos gramos enteros y no fracciones de gramos. Sin embargo, los valores subyacentes de peso, es decir, los valores exactos, sí admiten valores intermedios, por lo que una variable como esta será considerada continua. Esta situación es análoga a la discutida para el tiempo semanal de actividad física, que también corresponde a una variable cuantitativa continua.

Notemos que, si bien tanto variables cualitativas ordinales como cuantitativas discretas toman valores que son ordenables, en una variable cuantitativa discreta es posible cuantificar la distancia entre dos valores posibles. Sabemos que la distancia entre tener **2** hermanos y **5** hermanos son **3** hermanos. Sin embargo, esto no es cierto para las variables cualitativas ordinales; como lo vemos, por ejemplo, en el caso de los niveles socioeconómicos: no existe una manera de cuantificar la diferencia entre los niveles socioeconómicos ABC1, C2, C3, D y E. Solo podemos dar un orden a estos valores.

Muchas veces, los valores de variables cualitativas son codificados al momento de realizar la encuesta o cuando se ingresan las respuestas para ser analizadas. A modo de ejemplo, en una encuesta sobre los colores favoritos de los niños, los colores pueden ser registrados como rojo: **1**, azul: **2**, verde: **3**, etc. Sin embargo, los valores asignados carecen de significación, puesto que los colores no son en realidad cuantificables, ni menos ordenables. De este modo, los valores **1**,

2, 3, etc. corresponden a códigos que no deben ser tratados como numéricos, ni menos realizar operaciones aritméticas con ellos. Lo mismo ocurre si consideramos los números de las camisetas de los jugadores de un equipo de fútbol. Estos números solo son utilizados como identificadores y no como cuantificadores. De este modo, se debe tener cuidado con no identificar a una variable como cuantitativa solo porque está formada por dígitos.

Como veremos en el capítulo siguiente, la clasificación de una variable, ya sea como cualitativa, cuantitativa y sus alternativas, determina el tipo de representación y análisis adecuado que haremos de ella, al disponer de los datos de la muestra.

En resumen

- Una *variable cualitativa* es aquella cuyos valores posibles corresponden a categorías.
- Una variable cualitativa se dice *cualitativa ordinal*, si existe un orden natural para sus categorías. En caso contrario, se dice que la variable es *cualitativa nominal*.
- Una *variable cuantitativa* es aquella que toma valores numéricos.
- Una variable cuantitativa se dice *cuantitativa continua* si, entre dos valores cualesquiera de ella, siempre existe otro valor posible. En caso contrario, se dice que la variable es *cuantitativa discreta*.

Ejercicios

1. En cada una de las siguientes situaciones, identifique la variable a medir y su tipo.
   a. En cierta municipalidad se desea estudiar el ingreso per cápita de los habitantes de la comuna.
   b. Se está interesado en medir el grado de aprobación de la población mayor de 21 años a las medidas educacionales planteadas por el gobierno, para lo cual se plantea la afirmación: "El aumento de horas de Lenguaje en las escuelas mejorará los puntajes obtenidos en las pruebas SIMCE". Las personas encuestadas deben elegir una de cinco alternativas, entre:

      Muy en desacuerdo – En desacuerdo – Ni de acuerdo ni en desacuerdo – De acuerdo – Muy de acuerdo

   c. En una escuela se desea investigar la cantidad de hermanos que tiene cada alumno.
   d. Se desea averiguar el porcentaje de hogares del país que han sido víctimas de asalto.
2. Ana y Luisa están recolectando datos sobre el tiempo que toma a los participantes de una maratón de 5 kilómetros completar la carrera. Ana dice que la variable es continua, ya que los tiempos de competición pueden tomar cualquier valor entre 0 segundos y 2 horas (cuando la maratón termina oficialmente). Luisa piensa que la variable es discreta, ya que contamos horas, minutos y segundos. ¿Con quién está de acuerdo? Explique su razonamiento.

Ejercicios del capítulo

1. En cierto estudio, se desea conocer las preferencias de los electores sobre los candidatos en una elección presidencial, con el objetivo de realizar predicciones sobre el resultado de la elección. Elija cuál sería la mejor definición de la población de interés en este caso:
   a. Todos los individuos que viven en el país.
   b. Todos los individuos chilenos inscritos en algún partido político.
   c. Todos los individuos chilenos mayores de 18 años.
   d. Todos los individuos chilenos residentes en la Región Metropolitana.

2. En la situación anterior, ¿qué subpoblaciones se podrían comparar?

3. Alumnos de sexto básico desean realizar una actividad para recaudar fondos para la compra de varios juegos de mesa para donar a la escuela. Ellos deciden realizar una encuesta y así recolectar información sobre la preferencia de juegos de mesa de todos los niños de la escuela. Los encuestados fueron un grupo de alumnos sentados en la cafetería.
   a. ¿Cuál es la población en este contexto?
   b. Describa las limitaciones de la muestra.
   c. Explique una mejor manera de obtener una muestra representativa.

4. En cierto estudio, se decidió encuestar a 500 personas seleccionadas de manera aleatoria en 15 localidades de 5 países, elegidos de tal manera que todos los países sudamericanos de habla hispana tuvieron las mismas oportunidades de pertenecer a la muestra. Si se le indica que esta muestra es representativa, ¿cuál cree usted que es la población de interés?
   a. Todas las personas en países sudamericanos de habla hispana.
   b. Solo las personas de las 15 localidades seleccionadas.
   c. Las personas de los 5 países sudamericanos donde se realizó la encuesta.
   d. Las 500 personas encuestadas.

5. Considere la pregunta: "¿Debe el Congreso continuar estando en Valparaíso?" Un canal de televisión pidió a los televidentes que llamaran para dar su opinión al respecto. De 186.000 televidentes que llamaron, un 67% respondió negativamente a la pregunta. Por otra parte, una encuesta a nivel nacional seleccionó una muestra aleatoria simple de solo 500 adultos, y se encontró que un 72% respondió afirmativamente a la pregunta. Explique por qué las opiniones de estas 500 personas son una mejor guía del pensamiento del país, que las opiniones de las 186.000 personas que llamaron al programa televisivo.

6. En cada una de las siguientes situaciones, identifique la muestra utilizada:
   a. En una escuela se desea investigar el número de hermanos que posee cada niño. Para esto, se pregunta cuántos hermanos tienen a un grupo de 20 niños de la escuela.
   b. Se desea investigar el porcentaje de hogares en el país que ha sido víctima de un asalto durante el último año. Para esto, se entrevistan 100 hogares elegidos en 15 comunas diferentes.

7. En las siguientes situaciones, indique si el tipo de muestreo utilizado es probabilístico. En caso de serlo, identifique si, en alguna de sus partes, se realizó un muestreo aleatorio simple.

   a. Por cada **10** personas que salen del estadio, se le pregunta a una qué le pareció el evento.

   b. Para un estudio sobre obesidad escolar, se seleccionan **50** participantes entre primero y sexto básico, **50** participantes entre quinto y octavo básico, y **50** participantes entre primero y cuarto medio.

8. En los estudios que utilizan encuestas, a menudo se hacen declaraciones como las que se muestran a continuación. Analice si son correctas o incorrectas.

   a. Siempre es mejor tomar un censo que seleccionar una muestra.

   b. Una buena manera de muestrear, si deseamos saber acerca de la calidad de comida en la cafetería, es detener a alumnos que van camino a esta.

   c. Una encuesta tomada de un sitio de Internet que da apoyo a la enseñanza de la estadística recolectó **12.357** respuestas. La mayoría de quienes respondieron a la encuesta expresó que disfrutaba hacer tareas de estadística. Debido al gran tamaño muestral, podemos estar seguros de que la mayoría de alumnos de estadística opina igual.

9. Considere una población de **6** niños, donde las edades de cada uno son:

Niño	Efraín	Eduardo	Catalina	Ailén	Elisa	Alfredo
Edad (años)	10	12	9	10	11	13

   a. Obtenga el promedio de las edades de los **6** niños. Esta cantidad, ¿corresponde a un parámetro o a un estadístico?

   b. Considere todas las muestras posibles formadas por **2** de estos niños y, para cada una de ellas, obtenga el promedio de las edades. Estas cantidades, ¿corresponden a parámetros o estadísticos?

   c. Compare el valor del promedio de las edades de los **6** niños, con los promedios de edades obtenidos en cada una de las muestras de **2** niños. ¿Qué observa?

10. Sus alumnos han mostrado interés por estudiar si existen diferencias entre los hábitos del estudio de su curso, el quinto A y el curso paralelo, el quinto B.

    a. ¿Qué variables cree usted que sería interesante que ellos recolectaran? Proponga al menos **5**.

    b. Para cada una de las variables propuestas en a., indique si ella es cualitativa, nominal u ordinal, o cuantitativa, discreta o continua.

    c. ¿Qué preguntas podría formular a sus alumnos para motivar sus propias propuestas?

11. Con el objeto de determinar los factores que pueden influir en el rendimiento del equipo de fútbol de la escuela en el campeonato interescolar, Andrea decide recolectar información tanto de su equipo, como de los equipos rivales.

    a. Mencione **5** variables relacionadas a cada equipo que considere de interés para los objetivos de Andrea. Explique por qué lo son.

b. Para cada una de las 5 variables mencionadas en el apartado anterior, indique si ella es cualitativa, nominal u ordinal, o cuantitativa, discreta o continua.

12. Para cada una de las siguientes variables, indique si ella es cualitativa, nominal u ordinal, o cuantitativa, discreta o continua.

a. Número de visitas al doctor de Cristina en un año.
b. Nivel de escolaridad: solo Educación Básica – hasta Educación Media – Educación Superior.
c. Cantante favorito de las mamás de los alumnos del cuarto básico.
d. Tiempo de espera hasta que pasa el autobús en un paradero.
e. Cantidad de hojas en el cuaderno de María José.
f. Altura del escritorio de la profesora García.
g. Número de hamburguesas encargadas para la celebración del cumpleaños de Felipe.
h. Colores de las alianzas para la semana de la escuela.

Capítulo
III

Organización de datos y representación de la información

Introducción

Como estudiamos anteriormente, en las etapas iniciales de un ciclo de investigación se plantea la o las preguntas de interés, se determina la información que se requerirá para responderlas, y se recolectan los datos necesarios para extraer esta información. En este capítulo, se presentan formas de organizar los datos obtenidos, de manera de extraer la mayor cantidad de información relevante a partir de estos. Las herramientas que estudiaremos son válidas para conjuntos de observaciones en general, por lo que no haremos distinciones sobre cómo estas se obtuvieron, pudiendo corresponder tanto a una muestra como a una población completa.

Este capítulo está organizado como sigue: en la **Sección 1** abordaremos la importancia de las representaciones que se estudiarán y la necesidad de enseñar su construcción e interpretación a niños en edad escolar. Las **Secciones 2** y **3** constituyen el centro de este capítulo y se concentran en dos tipos complementarios de representación de datos. En la **Sección 2** trataremos la representación de datos a través de tablas de frecuencias, donde estudiaremos representaciones para una variable de manera individual, extendiéndolo luego a representaciones para estudiar el comportamiento de dos variables de manera conjunta. Por otra parte, en la **Sección 3** trataremos representaciones gráficas de los datos, donde distinguiremos desde gráficos concretos y reales, hasta representaciones más abstractas, como gráficos de barras, histogramas y gráficos circulares, entre otras. Finalmente, en la **Sección 4** discutiremos estos dos tipos de representaciones con una visión general, destacando su carácter complementario.

1. Motivación

En la actualidad, nos vemos constantemente enfrentados a información entregada a través de gráficos y tablas, tanto en medios de comunicación o en situaciones de tipo cotidiano, como en publicaciones o estudios más específicos sobre temas de nuestro interés personal, o relacionados con nuestro quehacer laboral. Esta información es, por lo general, el resultado de estudios originados por una pregunta o problema de interés. Todo ciudadano debe estar preparado para extraer correctamente la información entregada a través de estas fuentes, de modo de aprender críticamente sobre su entorno y ser capaz de tomar decisiones. Un estudiante debe, además, conocer y entender qué representaciones resultan más apropiadas para contestar o abordar las preguntas de interés en el ciclo de investigación.

A modo de ejemplo, las **Figuras III.1** y **III.2** muestran información en forma de gráficos y tablas encontrados en diferentes medios de difusión. A la izquierda [1], la **Figura III.1** entrega información sobre el uso de Internet en niños y jóvenes en edad escolar, como parte de un estudio orientado a medir el grado de digitalización (acceso, conocimiento, uso y valoración de Internet) de los escolares chilenos en centros urbanos del país, diferenciando por el tipo de dependencia del establecimiento educacional al que asisten. A la derecha [2], la **Figura III.1** entrega información sobre la evolución, entre los años 2005 y 2011, de la matrícula en programas de Educación Superior relacionados a la minería, como parte de un estudio para proyectar la demanda de capital humano en esta área.

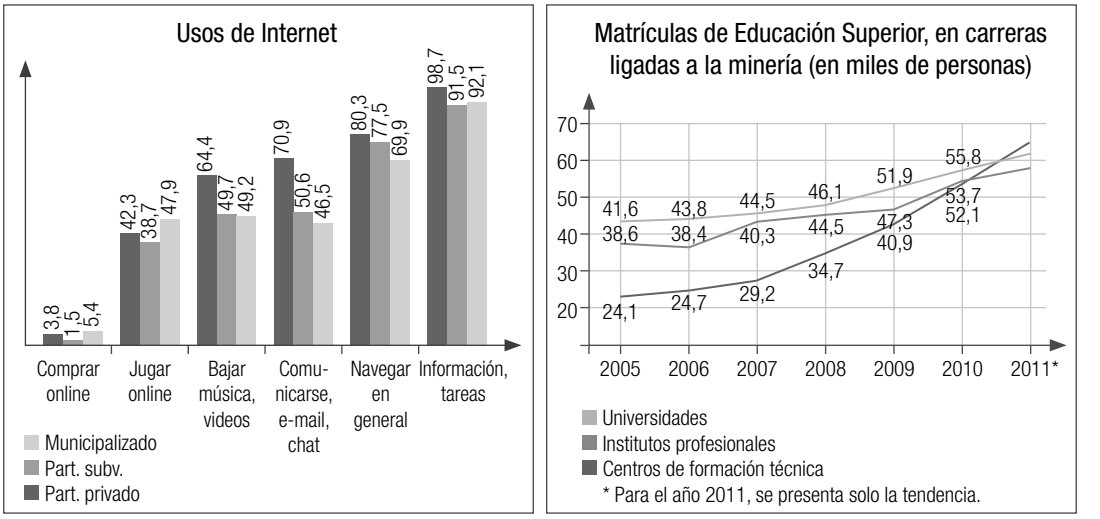


Figura III.1: Diferentes representaciones de datos en tablas y gráficos, encontrados en medios de difusión.

[1] Tomado del informe "Índice Generacional Digital", preparado por Educarchile, año 2005. http://beta.educarchile.cl/ech/pro/app/detalle?id=88168

[2] Tomado del informe "Fuerza laboral de la Gran Minería chilena 2012-2020", preparado por el Consejo de Competencias Mineras. http://www.consejominero.cl/wp-content/uploads/2013/04/Fuerza-Laboral-de-la-Gran-Mineria-Chilena-2012-2020.pdf

En la **Figura III.2**[3], la tabla muestra la distribución de la ocupación de las personas por rama de actividad económica y pertenencia a pueblos indígenas, como parte de un estudio para detectar la presencia de discriminación salarial en pueblos indígenas chilenos.

Distribución de los ocupados por rama de actividad económica
según pertenencia a pueblos indígenas y zona de residencia 2006

	Urbano		Rural	
Actividad	No indígena	Indígena	No indígena	Indígena
Act. No Bien Especificadas	1,2	0,5	0,4	0,3
Agricultura, Caza y Silvicultura	6,4	7,7	58,1	66,1
Explotación Minas y Canteras	1,8	1,0	1,3	0,6
Industrias Manufactureras	14,2	15,3	7,9	7,8
Electricidad, Gas y Agua	0,5	0,6	0,6	0,5
Construcción	9,7	12,4	5,3	5,5
Comercio Mayor/Menor	21,1	22,5	8,5	7,1
Transporte y Comunicaciones	8,2	7,4	3,7	1,4
Establecimientos Financieros	8,2	4,4	1,7	0,8
Servicios Comunales Sociales	28,8	28,2	12,5	9,9
Total	**100,0**	**100,0**	**100,0**	**100,0**

Fuente: Encuesta Casen 2006

Figura III.2: Diferentes representaciones de datos en tablas y gráficos, encontrados en medios de difusión.

Una de las formas para familiarizar a los alumnos con este tipo de herramientas de comunicación de información es enfrentándolos al ciclo de investigación. Los alumnos verán la necesidad de resumir los datos recolectados, en la búsqueda de la respuesta a una pregunta que ellos mismos, o bien sus profesores, hayan planteado. Para poder escoger las representaciones apropiadas, de acuerdo a las preguntas de interés, y para extraer información a partir de tablas y figuras como estas, es necesario conocer sus convenios de construcción, lo que hace necesaria su introducción en edad escolar temprana.

[3] Tomado del estudio "Segregación ocupacional y discriminación salarial en Chile: el caso de los pueblos indígenas", preparado por el Ministerio de Planificación, año 2009.
http://www.ministeriodesarrollosocial.gob.cl/btca/txtcompleto/mideplan/estudios_soc.1.pdf

2. Tablas de frecuencias

Para entender la utilidad de las tablas de frecuencias, consideremos la siguiente situación: en un curso de 15 alumnos, los niños están interesados en conocer sus deportes favoritos, por lo que cada uno de ellos indica cuál es el suyo. Dos alumnos del curso, Antonia y Diego, registran la información recolectada como se muestra en las **Tablas III.1** y **III.2**, de Antonia y Diego, respectivamente.

Nombre	Deporte favorito
Matías	Fútbol
Andrea	Tenis
María José	Gimnasia rítmica
Pablo	Fútbol
Antonia	Fútbol
Catalina	Gimnasia rítmica
Tomás	Básquetbol
Diego	Tenis
Lucas	Fútbol
Patricio	Básquetbol
Fernanda	Fútbol
Hernán	Básquetbol
Daniel	Básquetbol
Manuel	Fútbol
Eugenio	Gimnasia rítmica

Tabla III.1: Información registrada por Antonia.

Deporte favorito	Número de niños
Fútbol	6
Básquetbol	4
Gimnasia rítmica	3
Tenis	2

Tabla III.2: Información registrada por Diego.

Las tablas de Antonia y Diego son dos formas de registrar la información de interés, y la utilidad de cada una dependerá de lo que se desee comunicar.

Para pensar

¿En qué casos es preferible una de las tablas sobre la otra?

Notamos que la tabla de Antonia permite responder preguntas como: ¿cuál es el deporte favorito de María José? o ¿prefiere Andrea la gimnasia rítmica? Por otra parte, si consideramos preguntas como: ¿cuántos alumnos del curso tienen como deporte favorito el fútbol? o ¿qué deporte prefiere la mayoría de los niños?, aun cuando podemos extraer esta información a partir de la tabla de Antonia, resulta más fácil leerla directamente de la tabla de Diego.

De las observaciones anteriores notamos que, aunque una tabla como la de Diego entrega información menos detallada sobre los niños, a partir de ella es posible obtener descripciones generales del comportamiento de las observaciones. Así, por ejemplo, podremos contestar preguntas referidas al curso completo, y no a algunos alumnos en particular.

Supongamos, por ejemplo, que el objetivo de la recolección de los datos es ayudar a tomar la decisión sobre los deportes que se dictarán en los próximos talleres deportivos del curso. En ese caso, no nos interesa saber específicamente qué deporte prefiere cada niño, sino un resumen de la información de todo el curso. Lo mismo ocurriría, por ejemplo, si el objetivo de la recolección de los datos fuese saber si los niños, en general, prefieren deportes en equipos versus deportes individuales. Una tabla como la de Diego nos entrega información relevante en estos casos. Este tipo de tablas se denomina *tabla de frecuencias*, ya que permite observar el número de veces que se repite cada categoría.

En lo que sigue, aprenderemos a construir tablas de frecuencias para organizar la información en los datos.

2.1. Propósitos de las tablas de frecuencias

Si bien, como veremos en la siguiente sección, una tabla de frecuencias incluirá más información que la que se muestra en la **Tabla III.2**, de las consideraciones hechas ya podemos inferir algunas de las funciones que cumplen estas tablas. Todas estas funciones tienen como objetivo entregar luces sobre el comportamiento global de las observaciones.

El primer propósito o función de las tablas de frecuencias es resumir las observaciones. Es decir, representar de manera sucinta la información conjunta entregada por estas, eliminando información individual de cada una que pudiese dificultar la identificación de patrones globales.

El segundo propósito o función es organizar las observaciones. Podemos considerar que en ocasiones las tablas de frecuencias corresponden a un paso intermedio entre las observaciones y una representación gráfica, como veremos más adelante. En efecto, algunos gráficos suelen surgir fácilmente luego de haber organizado la información en tablas de frecuencias.

El tercer propósito o función es comunicar información. Este nace del objetivo que se planteó al recolectar las observaciones. Una tabla de frecuencias, en conjunto con otras herramientas, permite comunicar información recolectada, de manera de ayudar a responder la pregunta de investigación. Como mencionamos anteriormente, y veremos en las secciones que siguen, las tablas de frecuencias incorporan información que permite la extracción de conclusiones generales.

2.2. Tablas de frecuencias para una variable cualitativa

2.2.1 Frecuencias absolutas, relativas y relativas porcentuales

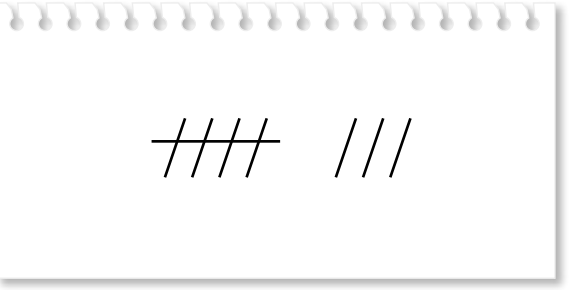
La construcción de una tabla de frecuencias se basa en el conteo de los datos en cada categoría de una variable. Una manera en que los niños en edad temprana puedan realizar esta labor corresponde a la construcción de lo que se denomina una *tabla de conteo* o *de registro*, como la Tabla III.3, que ha sido construida con los datos referentes a los deportes favoritos de los niños, de la sección anterior.

Deporte favorito	Registro
Fútbol	//////
Básquetbol	////
Gimnasia rítmica	///
Tenis	//

Tabla III.3: Tabla de conteo o registro construida con las respuestas de los niños.

Como vemos en el ejemplo, una tabla de conteo corresponde a una tabla de dos columnas, donde la primera de ellas indica las categorías de la variable de interés, en este caso, deporte favorito, y la segunda registra las observaciones en cada una de estas categorías, permitiendo así su "conteo". En esta segunda columna, la ocurrencia de cada caso se indica a través de una línea diagonal.

Una estrategia que los niños pueden utilizar para llenar la columna de conteos es que, cada vez que se obtenga una observación, se haga una marca en la categoría mencionada. Para facilitarles el conteo final, cada 4 marcas que se registren, es usual anotar la quinta como una línea horizontal, que tache las 4 primeras. A modo de ejemplo, si en una categoría existen 8 observaciones, los niños pueden realizar el registro como:



Con la información contenida en la tabla de conteo que se muestra arriba, los niños pueden responder algunas preguntas de interés; como, por ejemplo, ¿cuál es el deporte favorito de los niños? Si se han dibujado las marcas aproximadamente del mismo tamaño y con la misma separación, esta respuesta puede deducirse fácilmente identificando la fila más larga de marcas. También podemos responder ¿cuántos niños prefieren el básquetbol? Esto se debe a que la muestra contiene un número reducido de observaciones y los niños pueden contar rápidamente los

registros, lo que se conoce como *subitización*[4]. Si el número de observaciones en una categoría es muy grande, ya no será posible subitizar. En este caso, es conveniente cuantificar la información de la tabla de conteo pasando a una *tabla de frecuencias*. Para esto, los niños pueden realizar el conteo de los registros de las observaciones en cada categoría, y luego escribir los números correspondientes en una nueva columna de la tabla, como se muestra en la **Tabla III.4**, en la columna denominada *frecuencia absoluta*.

Deporte favorito	Registro	Frecuencia absoluta
Fútbol	//////	6
Básquetbol	////	4
Gimnasia rítmica	///	3
Tenis	//	2

Tabla III.4: Tabla de conteo o registro con columna para la frecuencia absoluta construida con las respuestas de los niños.

Dado que para responder las preguntas planteadas no es necesario rehacer el conteo, es posible eliminar la columna de registro, y obtener una tabla como la **Tabla III.5**.

Deporte favorito	Número de niños (Frecuencia absoluta)
Fútbol	6
Básquetbol	4
Gimnasia rítmica	3
Tenis	2
Total	15

Tabla III.5: Tabla de frecuencias absolutas construida con las respuestas de los niños.

El número de observaciones en cada categoría, en este caso, el número de niños, corresponde a la frecuencia absoluta de la categoría. Notamos que al sumar las frecuencias absolutas de todas las categorías se obtiene el valor **15**, es decir, el número total de observaciones, en este caso, el número total de niños en el curso, que se muestra en la celda inferior de la columna de frecuencias absolutas de la **Tabla III.5**. Formalmente, si anotamos la frecuencia absoluta de la categoría ***i*** por f_i, podemos expresar lo anterior como:

$$f_1 + f_2 + ... + f_I = n$$

Donde I corresponde al número de categorías de la variable de interés, y n al número total de observaciones. En el ejemplo, se tiene que $I = 4$ categorías. La categoría 1 corresponde a fútbol, la categoría 2 a básquetbol, la categoría 3 a gimnasia rítmica y la categoría 4 a tenis. De este modo, $f_1 = 6$, $f_2 = 4$, $f_3 = 3$ y $f_4 = 2$. Finalmente, en la expresión verificamos que:

$$f_1 + f_2 + f_3 + f_4 = 6 + 4 + 3 + 2 = 15$$

Que corresponde al número total de observaciones, n.

[4] Capacidad de enunciar rápidamente el número de objetos de una colección, por simple percepción global.

Las frecuencias absolutas nos permiten comparar las magnitudes de dos categorías del conjunto de datos. Por ejemplo, 6 niños prefieren el fútbol, mientras solo 2 prefieren el tenis, por lo cual existe una diferencia de 4 niños entre estas 2 categorías.

Para pensar

Una diferencia de 4 observaciones entre las frecuencias absolutas de dos categorías dadas, ¿tendrá la misma importancia si el total de las observaciones es 15, que si es 2.000?

Para las 15 observaciones, las categorías fútbol y tenis tienen frecuencias absolutas de 6 y 2, respectivamente. En este caso, la diferencia de 4 unidades en sus frecuencias nos puede parecer algo relevante. Por el contrario, si tuviésemos un conjunto de 2.000 observaciones, donde las categorías fútbol y tenis tuviesen frecuencias absolutas de 704 y 700, respectivamente, la misma diferencia de 4 unidades nos puede parecer algo irrelevante. Del mismo modo, si queremos comparar la frecuencia absoluta de una categoría con el total de observaciones, una frecuencia de 6 observaciones parece importante si el total es 15, pero no lo parece si el total de observaciones es 2.000.

En este sentido, las *frecuencias relativas* nos ayudan a entender de mejor manera la información. La frecuencia relativa corresponde a la proporción de observaciones de una categoría respecto del total de observaciones y, para calcularla, debemos dividir el número de observaciones de la categoría por el número total de observaciones, tal como se muestra en la **Tabla III.6.** De este modo, estamos estandarizando las frecuencias, de acuerdo al tamaño del conjunto de observaciones, lo que, además de permitirnos comparar adecuadamente la presencia de las categorías en un mismo conjunto, nos posibilita comparar conjuntos de datos diferentes, aun cuando sean de distinto tamaño.

Deporte favorito	Número de niños (Frecuencia absoluta)	Proporción de niños (Frecuencia relativa)
Fútbol	6	$\frac{6}{15} = 0{,}4$
Básquetbol	4	$\frac{6}{15} \approx 0{,}27$
Gimnasia rítmica	3	0,20
Tenis	2	0,13
Total	**15**	**1,00**

Tabla III.6: Tabla de frecuencias construida con las respuestas de los niños.

Por corresponder a una proporción, la frecuencia relativa puede ser representada por una fracción, un número decimal o un porcentaje. Así, por ejemplo, si observamos la

tercera columna de la **Tabla III.6**, la frecuencia relativa de ocurrencia de fútbol está representada por la fracción $\frac{6}{15}$, lo que significa que **6** de los **15** niños encuestados prefieren el fútbol. Por otra parte, si calculamos el número decimal o valor de la proporción, que corresponde a $\frac{6}{15} = 0{,}40$, esto se debiera interpretar como que **0,4** veces el total de observaciones corresponde al fútbol como deporte favorito. Dada la complejidad que pudiese tener la interpretación de una frecuencia relativa como decimal, es preferible interpretarla como porcentaje, es decir:

$$\frac{6}{15} \text{ x } 100\% = 40\%$$

Esto se interpreta como que el **40%** del total de observaciones, en este caso, de niños, corresponde al fútbol como deporte favorito. Cuando la frecuencia relativa se expresa en porcentaje, se denomina *frecuencia relativa porcentual*.

Al obtener la frecuencia relativa de la categoría fútbol como número decimal, encontramos que este valor, **0,4**, tiene un número finito y pequeño de decimales. Sin embargo, esto no siempre es así, como, por ejemplo, en la categoría básquetbol de la misma **Tabla III.6**. En efecto, la frecuencia relativa de esta categoría corresponde a $\frac{4}{15}$, cuya representación decimal es $0{,}2\overline{6}$. En la tabla, este valor ha sido aproximado a **0,27**, utilizando un dígito significativo [5]. Si bien esta puede ser una sugerencia, el número de dígitos significativos utilizados dependerá de cada situación en particular. Todo esto también es válido al presentar las frecuencias relativas porcentuales.

Notemos que la suma de las frecuencias relativas en la tabla corresponde a **1**, lo que se muestra en la última fila de la columna de frecuencias relativas.

Formalmente, si r_i representa la frecuencia relativa de la categoría i, esta se calcula como:

$$r_i = \frac{f_i}{n}$$

En el ejemplo, para fútbol, $f_i = 6$, $n = 15$ y

$$r_i = \frac{6}{15} = 0{,}4$$

Con esta misma notación, encontramos que:

$$\begin{aligned} r_1 + r_2 + \cdots + r_I &= \frac{f_1}{n} + \frac{f_2}{n} + \cdots + \frac{f_I}{n} \\ &= \frac{f_1 + f_2 + \cdots + f_I}{n} \\ &= \frac{n}{n} = 1 \end{aligned}$$

Es decir, la suma de las frecuencias relativas siempre es **1**, lo que habíamos verificado en la **Tabla III.6**.

[5] Para una definición de dígitos significativos, ver "Capítulo 1 de libro *Geometría*" de esta misma colección.

Del mismo modo, la suma de las frecuencias relativas porcentuales es siempre 100%.

Notemos que las frecuencias relativas porcentuales son números entre 0 y 100%, de modo que categorías con una frecuencia relativa porcentual cercana a 0 son categorías que contienen pocas observaciones del total. Por el contrario, categorías con una frecuencia relativa porcentual cercana a 100%, son categorías que contienen muchas observaciones del total. A modo de ejemplo, la frecuencia relativa porcentual del tenis corresponde a, aproximadamente, 13% del total de observaciones, mientras que la del fútbol es de un 40% del total de observaciones, lo que concuerda con el hecho que para muchos más niños su deporte favorito es el fútbol y no el tenis.

Se debe tener cuidado con la introducción de frecuencias relativas y relativas porcentuales en los niños. Diversa literatura internacional indica que, en lo posible, se debe postergar el uso de estas, hasta que los niños hayan alcanzado un nivel aceptable de comprensión del razonamiento proporcional.

En resumen

- La *frecuencia absoluta* de una categoría corresponde al número de observaciones en dicha categoría.
- La *frecuencia relativa* de una categoría corresponde a la proporción de observaciones en dicha categoría, con respecto al total de observaciones, y se obtiene dividiendo la frecuencia absoluta por el número total de observaciones.
- La *frecuencia relativa porcentual* de una categoría corresponde al porcentaje de observaciones en dicha categoría, con respecto al total de las observaciones.

Ejercicios

1. Suponga que se obtiene una muestra de 50 apoderados de una escuela, y se les pregunta a cada uno de ellos su nivel de satisfacción con la gestión de la directiva del Centro de Padres. Las opiniones obtenidas se resumen en la siguiente tabla:

Nivel de satisfacción	Número de apoderados	Porcentaje de apoderados
Muy insatisfecho	1	
Insatisfecho	5	
Indiferente	15	
Satisfecho	20	
Muy satisfecho	9	

   a. Complete la tabla obteniendo los porcentajes de apoderados en cada una de las categorías. Redondee al entero más cercano.

b. ¿A qué columna corresponde la frecuencia absoluta? ¿A qué columna corresponde la frecuencia relativa porcentual?

c. Verifique que la suma de las frecuencias absolutas corresponda al número total de apoderados encuestados y que la suma de las frecuencias relativas porcentuales corresponda a 100%.

2. En el problema anterior, responda:

a. ¿Qué nivel de satisfacción es el más frecuente? Justifique.

b. ¿Qué porcentaje de apoderados declara estar insatisfecho con la gestión de la directiva?

c. Entregue dos conclusiones generales sobre el nivel de satisfacción de los apoderados en base a los valores de la tabla. Justifique.

3. Un hospital presentó un estudio con el fin de analizar el uso dado a las reuniones semanales del equipo médico y los alumnos en práctica. Para esto, se observó un total de 21 reuniones y se registraron los diferentes asuntos tratados. Luego, estos asuntos fueron clasificados en 7 temas, que se muestran en la tabla a continuación. La tabla muestra la frecuencia absoluta de cada uno de los temas que sugirió el equipo médico.

Tema	Número de asuntos tratados	Porcentaje
Rutinas del hospital	20	
Alumnos en práctica	4	
Pacientes	3	
Visitas a pacientes	2	
Actividades científicas	8	
Actividades sociales	11	
Estructura física	8	
Total		

a. Complete las celdas faltantes de la tabla. Redondee las frecuencias relativas porcentuales al decimal.

b. ¿Cuántos asuntos propuso el equipo médico en las 21 reuniones del hospital?

c. ¿Qué tema fue el predominante?

d. ¿A qué porcentaje de los asuntos tratados corresponde este tema?

e. ¿Qué diferencia porcentual existe entre los temas relacionados con las actividades científicas y las sociales? Comente.

f. ¿Qué porcentaje de los asuntos tratados se relaciona directamente con los pacientes?

g. ¿Se puede afirmar que se dedicó más tiempo a asuntos relacionados a rutinas del hospital que a actividades científicas? Explique.

h. Señale dos conclusiones generales que se puedan extraer de la tabla.

4. En la misma situación anterior, los asuntos propuestos por los alumnos en práctica se muestran en la siguiente tabla:

Tema	Número de asuntos tratados	Porcentaje
Rutinas del hospital	1	
Alumnos en práctica	2	
Pacientes	0	
Visitas a pacientes	0	
Actividades científicas	4	
Actividades sociales	1	
Estructura física	0	
Total		

a. Complete las celdas faltantes de la tabla. Redondee al primer decimal.
b. ¿Cuántos asuntos propusieron los alumnos en práctica durante todas las reuniones del hospital?
c. ¿Qué tema fue el predominante?
d. ¿A qué porcentaje de los asuntos tratados corresponde este tema?
e. ¿Qué diferencia porcentual existe entre los asuntos relacionados con las actividades científicas y las sociales? Comente.
f. ¿Qué porcentaje de los asuntos tratados se relaciona directamente con los pacientes?
g. Señale dos conclusiones generales que se puedan extraer de esta tabla. Compare estas conclusiones con las obtenidas para el equipo médico, en el problema anterior.

2.2.2 Comparación de diferentes conjuntos de observaciones en base a tablas de frecuencias

Como ya mencionamos, las frecuencias relativas o relativas porcentuales nos permiten comparar dos o más conjuntos de datos, aunque estos sean de diferentes tamaños. A modo de ejemplo, supongamos que en el problema sobre los deportes favoritos de los niños realizamos la pregunta separadamente a niños y niñas, de modo que nos interesa comparar las preferencias de deporte entre ambos sexos. Podemos pensar en niños y niñas como dos grupos de sujetos diferentes. Si los comparamos diciendo, por ejemplo, "solo 2 de las niñas prefieren el fútbol, mientras que en los niños este número es 4", información que se obtiene a partir de la **Tabla III.1**, daríamos la idea equivocada de que el fútbol es de mayor preferencia entre los niños que entre las niñas. Sin embargo, esto no es así, dado que el número de niñas y de niños en el curso es diferente. En efecto, el número de niñas en el curso es 5, mientras que el número de niños es 10. De este modo, los porcentajes de niñas y niños que prefieren el fútbol son ambos iguales a 40% (calculados a partir de $\frac{2}{5}$ para niñas, y $\frac{4}{10}$ para niños), que muestra que no existe relación entre el gusto por el fútbol y el sexo. Retomaremos este punto en la **Sección 2.4**, cuando estudiemos tablas de frecuencias para dos variables, o de doble entrada.

2.2.3 Respuestas en blanco

Es usual encontrar tablas de frecuencias que muestran el número de personas que no responden a una pregunta determinada, lo que también se denomina *número de respuestas faltantes*, o *de respuestas en blanco*. Consideremos, a modo de ejemplo, la siguiente tabla de frecuencias, **Tabla III.7**, entregada en el informe de la "5ta encuesta nacional de la juventud", realizada por el Instituto Nacional de la Juventud durante noviembre y diciembre de 2006[6].

Situación del entrevistado	Número de jóvenes	Porcentaje
Solo estudia	1.256.734	31,4%
Solo trabaja	1.120.307	28,0%
No estudia ni trabaja	509.557	12,7%
Solo buscando trabajo	462.761	11,6%
Buscando trabajo y estudiando	377.880	9,4%
Trabaja y estudia	258.027	6,5%
No responde	15.125	0,4%
Total	4.000.391	100%

Tabla III.7: Tabla de frecuencias de la situación educacional y laboral de los jóvenes.

La tabla entrega las frecuencias absolutas y relativas porcentuales (segunda y tercera columnas, respectivamente) de cada posible situación. De la tabla podemos leer, por ejemplo, que de un total de **4.000.391** jóvenes, **1.256.734** se encuentran únicamente dedicados al estudio, lo que corresponde a un **31,4%** de la población.

La **Tabla III.7** también indica que un **0,4%** de la muestra no responde, es decir, hay un **0,4%** de respuestas faltantes o en blanco. Resulta importante entregar esta información, puesto que pueden existir razones asociadas a la situación del entrevistado que lo lleven a no responder, como puede ser el caso, por ejemplo, de preguntas que se refieren a un tema sensible, en que la persona puede verse perjudicada al responder, o a un tema personal, donde la persona no se siente cómoda respondiendo. Contando con mayor información sobre los entrevistados, podríamos también intentar asociar la no respuesta a otras variables, como nivel socioeconómico, edad u otro.

En general, en el caso de que haya personas que no responden, se deben reportar los resultados como se ha hecho en la **Tabla III.7**, donde la categoría de las personas que no responden ha sido incluida en la tabla, y los porcentajes entregados han sido calculados con respecto a todas las personas entrevistadas. Por otra parte, se debe distinguir entre los casos en que la persona indica no saber, y las situaciones en que la persona decide no responder.

Como convención, la categoría de individuos que no responden se ubica en la última fila de la tabla, como se ha hecho en la **Tabla III.7**. El siguiente ejercicio ilustra la importancia de incluir las respuestas en blanco, o no respuestas, dentro de la tabla de frecuencias.

[6] Tomado de la página web del Instituto Nacional de la Juventud. http://www.oij.org/es_ES/publicacion/v-encuesta-nacional-de-juventud-de-chile

Ejercicio

Suponga que, como parte de un estudio llevado a cabo por el departamento de formación de una escuela, se les pregunta a los alumnos sobre la frecuencia en que han sido objeto de violencia física o sicológica por parte de sus compañeros. Las alternativas son: muchas veces, a veces y nunca. De **30** alumnos en el curso, **2** alumnos contestan "muchas veces", **6** alumnos contestan "a veces", **18** alumnos contestan "nunca", y **4** alumnos no responden.

Se pide construir la tabla de frecuencias para ordenar la información contenida en los datos. Discuta la importancia de incluir las respuestas en blanco en la tabla y la relación que estas pudiesen tener con el grado de violencia escolar en dichos niños.

2.2.4 Definición de las categorías utilizadas

El ejemplo en la **Tabla III.7** muestra que es importante tener una definición clara de las categorías. A simple vista, las categorías "No estudia ni trabaja" y "Solo buscando trabajo" se intersectan, en el sentido que un joven que solo busca trabajo necesariamente no estudia ni trabaja, por lo que podría clasificarse en las dos categorías al mismo tiempo. Sin embargo, esta encuesta está diseñada para que cada persona pertenezca a una sola categoría, lo que se refleja en que la suma de las frecuencias relativas porcentuales es igual a **100%**[7]. De aquí, deducimos que la categoría "No estudia ni trabaja", en esta encuesta, no incluye a aquellos jóvenes que se encuentran buscando trabajo. Lo más adecuado hubiese sido renombrar las categorías de modo que representen de mejor manera lo que ellas significan.

En general, la convención al construir las categorías de una variable es que cada individuo o elemento debe pertenecer a una y solo una de estas categorías. Esto debe reflejarse claramente al asignar nombres a las categorías definidas. En el ejemplo que seguimos sobre la situación de los jóvenes, la categoría "No estudia ni trabaja" pudiese ser renombrada como "No estudia, no trabaja y no busca trabajo".

[7] Si una persona puede pertenecer a dos o más categorías simultáneamente, la suma de frecuencias relativas porcentuales sería mayor a 100%. Esto no corresponde a un error, sino a una metodología diferente de medición. Existen preguntas donde se pide al entrevistado elegir una o más alternativas que lo representen. En estos casos, no se debe presentar la fila inferior de la tabla que indica la suma de las frecuencias relativas porcentuales, porque podrían inducir a confusión.

2.2.5 Tablas de frecuencias para una variable cualitativa ordinal

En las tablas que hemos construido, la variable de interés (deporte favorito o situación laboral de los jóvenes) corresponde a una cualitativa nominal, por lo que el orden de las categorías en la tabla de frecuencias es irrelevante. Sin embargo, esto no es así al tratarse de variables cualitativas ordinales. Consideremos, a modo de ejemplo, la información entregada por la misma encuesta juvenil, donde se expresa la asistencia de los jóvenes a ceremonias religiosas. El informe entrega las frecuencias relativas porcentuales, como se muestra en la **Tabla III.8**.

Asistencia a ceremonias religiosas	Porcentaje de jóvenes
Nunca	44,1%
Solo ocasionalmente	33,7%
Semanalmente	13,4%
Una vez al mes	6,7%
No responde	2,1%

Tabla III.8: Frecuencias relativas porcentuales de la asistencia de los jóvenes a ceremonias religiosas (excluye bautizos, matrimonios y funerales).

Para pensar

En la **Tabla III.8**, ¿cuál cree que ha sido el criterio para ordenar las categorías? ¿este le parece razonable?

La variable de interés corresponde a una variable cualitativa ordinal, ya que sus categorías son ordenables: de menor a mayor asistencia a ceremonias religiosas ("Nunca", "Solo ocasionalmente", "Una vez al mes", "Semanalmente") o viceversa. Al parecer, el criterio para el ordenamiento de las categorías en la **Tabla III.8** ha sido el orden decreciente de las frecuencias relativas porcentuales, desde **44,1%** a **2,1%**. Sin embargo, mostraremos la conveniencia de utilizar, en vez, uno de los dos ordenamientos naturales de las categorías de la variable: de menor a mayor asistencia a ceremonias religiosas o viceversa.

Supongamos que nos interesa leer, por ejemplo, el porcentaje de jóvenes que no asiste o asiste a ceremonias religiosas una o menos veces al mes. Para esto, debemos sumar los porcentajes de las categorías "Nunca", "Solo ocasionalmente" y "Una vez al mes", obteniendo **(44,1 + 33,7% + 6,7)% = 84,5%**. La identificación de las categorías a sumar se dificulta, puesto que se encuentran separadas en la tabla. Resultaría más fácil leer esta información a partir de la tabla, si esta se hubiese construido desde menor a mayor asistencia.

En efecto, si se hubiesen ordenado las categorías en orden creciente de asistencia, sería posible obtener el porcentaje deseado sumando las frecuencias de las categorías "Una vez al mes" y anteriores. Esta cantidad se denomina *frecuencia relativa* (o *relativa porcentual*) *acumulada*. A modo de ejemplo, en el caso de la asistencia a ceremonias religiosas, podríamos entregar una tabla como la **Tabla III.9**, donde la última columna muestra la frecuencia relativa porcentual acumulada, que se obtiene sumando las frecuencias relativas de interés y anteriores.

Asistencia	Frecuencia relativa porcentual	Frecuencia relativa porcentual acumulada
Nunca	44,1%	44,1%
Solo ocasionalmente	33,7%	44,1% + 33,7% = 77,8%
Una vez al mes	6,7%	77,8% + 6,7% = 84,5%
Semanalmente	13,4%	97,9%
No responde	2,1%	100,0%

Tabla III.9: Tabla de frecuencias de la asistencia de los jóvenes a ceremonias religiosas (excluye bautizos, matrimonios y funerales). El orden de las categorías en la tabla sigue un ordenamiento natural, de menor a mayor asistencia.

A modo de ejemplo, la frecuencia relativa porcentual acumulada de la categoría "Solo ocasionalmente" se obtiene sumando las frecuencias relativas porcentuales **44,1% + 33,7% = 77,8%**, lo que se interpreta como que un **77,8%** de los jóvenes no asiste, o asiste solo ocasional mente, a ceremonias religiosas. Por otra parte, la frecuencia relativa porcentual acumulada de la categoría "Semanalmente" se obtiene sumando las frecuencias relativas porcentuales **44,1% + 33,7% + 6,7% + 13,4% = 97,9%**.

Volviendo a la pregunta de interés sobre el porcentaje de jóvenes que no asiste o asiste a ceremonias religiosas una o menos veces al mes, de la columna de frecuencias relativas porcentuales acumuladas en la **Tabla III.9**, se lee directamente la respuesta que dimos anteriormente: un **84,5%** de los jóvenes no asiste a ceremonias religiosas o asiste con una periodicidad de, a lo más, una vez al mes.

Como ya lo mencionamos, también podríamos presentar la tabla con las categorías ordenadas de mayor a menor asistencia a ceremonias religiosas. De este modo, las frecuencias relativas porcentuales acumuladas permitirían determinar, por ejemplo, el porcentaje de jóvenes que asiste al menos una vez al mes a una ceremonia religiosa. El orden utilizado dependerá de la información que se desee transmitir, sin que esto signifique que se está distorsionando la información. De cualquier modo, siempre es posible construir una tabla a partir de la otra.

Así, como en cualquier tabla de frecuencias, se acostumbra mostrar la categoría "No responde" al final de la tabla.

En resumen

- Si la variable considerada es de tipo cualitativa ordinal, se debe respetar el orden de sus categorías dentro de la tabla.
- En este caso, la *frecuencia relativa (*o *relativa porcentual) acumulada* de una categoría corresponde a la suma de las frecuencias relativas (o relativas porcentuales) de la categoría de interés y anteriores.

Ejercicios

1. Supongamos que se obtiene una muestra de **50** apoderados de una escuela. Se pregunta a cada uno de ellos su nivel de satisfacción con la gestión de la directiva del Centro de Padres. Las opiniones obtenidas se resumen en la siguiente tabla:

Nivel de satisfacción	Número de apoderados	Porcentaje de apoderados
Muy insatisfecho	1	
Insatisfecho	5	
Indiferente	15	
Satisfecho	20	
Muy satisfecho	9	

   a. Explique por qué en este caso tiene sentido obtener las frecuencias relativas acumuladas.

   b. Complete la tabla agregando a la derecha la columna de frecuencias relativas porcentuales acumuladas. Verifique que el último valor obtenido sea **100%**.

   c. ¿Qué porcentaje de los apoderados encuestados se declara "Insatisfecho" o "Muy insatisfecho" con la gestión de la directiva?

   d. ¿Qué porcentaje de los apoderados encuestados no está satisfecho con la gestión de la directiva?

   e. ¿Cómo ordenaría las categorías de la tabla, si quisiera leer directamente de ella el porcentaje de apoderados encuestados que está satisfecho con la gestión de la directiva?

   f. Si usted quisiera destacar que la mayoría de los apoderados encuestados está satisfecho con la gestión de la directiva, ¿qué ordenamiento de las categorías elegiría? Explique.

   g. Según lo pedido en el apartado anterior, notamos que el ordenamiento induce a una interpretación determinada. Discuta el mal uso que se puede dar a esto.

2. En cada uno de los siguientes casos, determine si es adecuado presentar frecuencias relativas, o relativas porcentuales, acumuladas en sus tablas de frecuencias. Justifique.

   a. En un estudio sobre el estado civil en un grupo de asistentes a una charla, se registraron **167** personas solteras, **27** personas casadas, **34** personas divorciadas, **7** personas viudas y **65** personas en unión libre.

   b. En un estudio sobre el nivel educacional de apoderados de una escuela, se registraron **56** apoderados que solo habían recibido Educación Básica, **167** apoderados que habían recibido hasta Educación Media, **42** apoderados que habían recibido Educación Universitaria completa y **13** apoderados que habían recibido especialización. Finalmente, **4** apoderados no respondieron.

c. Al construir una evaluación escolar de Lenguaje, el grupo de profesores del ciclo se reunió para evaluar su nivel de dificultad. Se encontró que 7 ítems correspondían a preguntas muy difíciles, 6 ítems a preguntas difíciles, 9 ítems a preguntas algo difíciles, y 1 pregunta poco difícil.

d. Un estudio de salud pública realizó una encuesta a una muestra de 157 pacientes que padecían presión alta y registró que un 52,8% fumaba, un 40,1% no fumaba y un 7,1% había dejado de fumar.

3. a. En cada uno de los casos anteriores que corresponda, construya una tabla de frecuencias que incluya las frecuencias relativas porcentuales acumuladas. Elija un orden adecuado de las categorías. Justifique.

b. De acuerdo a las tablas anteriores y el ordenamiento decidido, en cada una de ellas determine dos preguntas que puedan ser respondidas directamente de la tabla o con cierta integración de sus contenidos.

c. En cada uno de los casos anteriores, obtenga la tabla con el ordenamiento inverso al que utilizó en el apartado anterior. ¿Qué preguntas puede responder ahora?

2.2.6 Redondeo de decimales en las frecuencias relativas

Habíamos notado anteriormente que la suma de las frecuencias relativas porcentuales siempre debe ser igual a 100% (e igual a 1 en el caso de las frecuencias relativas). Sin embargo, debido a que nos vemos en la necesidad de redondear las cantidades para presentarlas en la tabla, es posible que esto no ocurra. Existen diferentes maneras de tratar esta situación.

Una primera estrategia corresponde a comenzar redondeando las frecuencias relativas por aproximación al decimal deseado y verificar luego que ellas sumen 100%. A modo de ejemplo, supongamos que en una encuesta a 110 personas sobre preferencias de ingredientes de pizza se obtuvieron los resultados que se muestran en la Tabla III.10, donde las frecuencias relativas porcentuales han sido redondeadas, por aproximación al entero más cercano.

Ingrediente	Número de entrevistados	Porcentaje
Anchoas	8	7%
Queso	27	25%
Peperoni	16	15%
Vienesas	36	33%
Vegetales	23	21%
Total	110	101%

Tabla III.10: Tabla de frecuencias sobre preferencia de ingredientes de pizza. Las frecuencias relativas porcentuales aproximadas al entero no suman 100%.

En la tabla notamos que la suma de las frecuencias relativas porcentuales corresponde a 101%, es decir 1 punto porcentual mayor a lo debido. Para subsanar esta situación, sugerimos dos estrategias alternativas: la primera es restar 1 punto porcentual a la frecuencia relativa porcentual

de la categoría más frecuente en la tabla. De este modo, el efecto de este punto será proporcionalmente menor que en las categorías restantes. En el ejemplo, disminuiremos en 1 punto porcentual la frecuencia relativa porcentual del ingrediente "Vienesas", cambiándola de 33% a 32%. La Tabla III.11 muestra las nuevas frecuencias relativas.

Ingrediente	Número de entrevistados	Porcentaje
Anchoas	8	7%
Queso	27	25%
Peperoni	16	15%
Vienesas	36	32%
Vegetales	23	21%
Total	110	100%

Tabla III.11: Tabla de frecuencias sobre preferencia de ingredientes de pizza. Las frecuencias relativas porcentuales suman 100%.

Por otra parte, en algunas ocasiones es posible lograr que la suma de las frecuencias relativas porcentuales que se muestran en la tabla sea exactamente 100%, utilizando un mayor número de decimales. Sin embargo, un número muy grande de decimales muchas veces dificulta la lectura.

Una buena alternativa corresponde a incorporar una nota al pie de la página, que indique que las cantidades de la tabla no suman 100% debido a redondeo. Esta alternativa parece ser la más simple y transparente. En este caso, no se indica la suma de las frecuencias relativas porcentuales en la última fila de la tabla, para no confundir al lector.

2.3. Tablas de frecuencias para una variable cuantitativa

2.3.1 Clases o categorías

Si bien las tablas de frecuencias que hemos presentado corresponden a tablas para variables cualitativas, también es posible utilizarlas para representar el comportamiento de variables cuantitativas.

A modo de ejemplo, la "5ta encuesta nacional de la juventud" estudió también la edad de los jóvenes, medida en años, que corresponde a una variable cuantitativa. Para hacer un paralelo con las tablas de frecuencias para variables cualitativas que hemos estudiado, podemos proponer que cada número de años en la muestra corresponda a una categoría. Sin embargo, si consideramos que las edades de los jóvenes van desde los 15 a los 29 años, el número de categorías sería muy grande e impediría visualizar de buena forma la distribución de la edad de los jóvenes, por ejemplo, detectando patrones en el comportamiento de estas.

En casos como el anterior, cuando la variable toma muchos valores, es usual agruparlos en intervalos, donde cada intervalo de valores corresponde a una categoría. Estas categorías son también, usualmente, denominadas *clases*.

La encuesta mencionada, por ejemplo, reporta sus resultados según se muestra en la Tabla III.12:

Tramos de edad	Número de jóvenes	Porcentajes
15 - 19 años	1.462.862	36,5%
20 - 24 años	1.343.012	33,6%
25 - 29 años	1.194.518	29,9%
Total	4.000.392	100%

Tabla III.12: Frecuencias absoluta y relativa porcentual de la edad de la población joven.

En la tabla, cada categoría o clase representa un rango o tramo de edad, por lo que se han definido 3 clases. Al igual que con variables cualitativas, es necesario que las clases sean excluyentes, es decir, en 2 clases no puede haber valores de la variable en común. En el caso de la encuesta, no se puede contar un mismo joven en dos clases diferentes. Por tratarse de variables cuantitativas, las categorías son ordenables y se debe respetar el orden en la tabla. En general, se utiliza el orden creciente de los valores de la variable, como en la Tabla III.12. Al igual que para variables cualitativas ordinales, en la tabla es posible entregar, además, las frecuencias relativas o relativas porcentuales acumuladas, aunque no se muestran en la tabla entregada por el estudio.

A modo de ejemplo, a continuación construiremos una tabla de frecuencias para la edad de los profesores de una escuela, y extraeremos información relevante de ella.

Suponga que las edades de los profesores de la escuela son:

32	37	36	32	51	53	33	61	35	45	55	39
76	37	42	40	32	60	38	56	48	48	40	43
62	43	42	44	41	56	39	46	31	47		

Debemos decidir cuáles serán las clases a considerar. Como veremos más adelante, es aconsejable que todas las clases sean del mismo tamaño. En este ejercicio, entenderemos tamaño como el número de valores posibles, edades en este caso, en cada una de las clases (sin importar si dicho valor pertenece o no al conjunto de observaciones con que se está trabajando). Formalmente, el tamaño se representará a través del concepto de *amplitud de la clase*, y nos referiremos a él en breve.

Por ahora, tomaremos clases que contengan 10 valores cada una. Dado que el menor valor es 32 y el mayor 76, decidimos que la primera clase irá desde los 30 hasta los 39 años, la segunda clase desde los 40 hasta los 49 años, y así sucesivamente, hasta la quinta clase que contendrá a los profesores entre 70 y 79 años.

El siguiente paso es obtener el número de profesores en cada categoría o clase, y luego obtener las frecuencias relativas o relativas porcentuales, como es habitual. En particular, la **Tabla III.13** resume los datos a través de las frecuencias absoluta, relativa porcentual y relativa porcentual acumulada.

Categoría o clase	Número de profesores	Porcentaje	Porcentaje acumulado
30 - 39 años	12	35%	35%
40 - 49 años	13	38%	73%
50 - 59 años	5	15%	88%
60 - 69 años	3	9%	97%
70 - 79 años	1	3%	100%
Total	34	100%	

Tabla III.13: Tabla de frecuencias de las edades de los profesores.

Una vez construida la tabla, podemos ver que el mayor porcentaje de profesores (38%) tiene entre 40 y 49 años. A partir de la columna de frecuencia relativa porcentual acumulada, leemos que la mayoría de los profesores (73%) es menor de 50 años. Según la columna de frecuencias relativas porcentuales acumuladas, podemos señalar que un 97% – 73% = 24% de los profesores tiene entre 50 y 69 años. A partir de una mirada global de la tabla, es posible concluir que el número de profesores disminuye con la edad de los mismos, ¿puede usted aventurar explicaciones?

Ejercicios

1. Las siguientes observaciones corresponden al promedio entre el máximo y el mínimo de las temperaturas diarias (en grados Celsius) observadas en una ciudad durante 50 días.

–1	0	0	1	1	2	2	3	3	3	4	4
5	5	6	6	6	6	7	7	8	8	8	8
9	9	9	11	11	12	13	13	13	13	13	13
14	15	16	16	17	17	18	18	18	18	19	19
21	26										

   a. Construya una tabla de frecuencias utilizando categorías que contengan 10 valores posibles (en grados Celsius) cada una, comenzando en –5. Incluya las frecuencias absolutas, relativa porcentual y relativa porcentual acumulada.
   b. ¿En qué categoría o clase de temperaturas se encuentra el mayor porcentaje de los días?
   c. ¿En qué porcentaje de los días se observa, a lo más, 14 grados?
   d. Obtenga una conclusión general del comportamiento de la temperatura en los días de invierno.

2. En el ejercicio anterior, obtenga una nueva tabla de frecuencias, definiendo categorías o clases que contengan 5 valores posibles (en grados Celsius). Señale una conclusión general del comportamiento de la temperatura en los datos. ¿Coincide con la conclusión obtenida en el apartado d. del ejercicio anterior?

3. En los ejercicios anteriores, ¿qué ventajas y desventajas encuentra en utilizar 5 y 10 valores en cada categoría?

2.3.2 Amplitud de la clase

Como dijimos, la amplitud de la clase es una manera de medir su tamaño. En general, cuando tratamos variables que toman valores enteros, es razonable representar su tamaño a través del número de valores posibles que hay dentro del intervalo. Sin embargo, este concepto no puede ser utilizado al tratar con variables continuas, dado que, por definición, ellas pueden tomar infinitos valores.

Consideremos, a modo de ejemplo, los siguientes pesos, medidos en kilogramos, de 10 alumnos de un curso:

40,2 40,5 38,3 42,0 44,1 43,2 46,7 45,5 42,3 47,4

Los pesos varían entre 38,3 y 47,4 kilogramos. Una propuesta podría ser utilizar las categorías: desde 38 a 40 kilos, sin incluir este último valor, desde 40 a 42 kilos, sin incluir este último valor, y así sucesivamente, hasta la categoría desde 46 a 48 kilos, sin incluir este último valor. ¿Cuál sería la amplitud de cada clase en este caso? Si consideramos la primera categoría, que parte en 38 kilogramos, ¿cuántos son los valores posibles dentro de ella? A primera vista, nos parece que los valores posibles son:

38,0 38,1 38,2, …, 39,7 39,8 y 39,9

Es decir, 20 valores. Según esto, la amplitud de la clase sería 20. ¿Qué ocurriría si nuestra balanza tuviera mayor precisión, de modo que pudiese también medir décimas de gramo? En este caso, los valores posibles de la primera categoría serían:

38,0 38,01 38,02, …, 39,97 39,98 y 39,99

Es decir, 200 valores, por lo que la amplitud de la clase sería 200. Esta ambigüedad surge ya que la variable que estamos considerando es continua. En la medida que nuestro instrumento de medición sea más y más preciso, aparecerán más valores posibles para las observaciones, lo que se debe a que la variable que estamos midiendo, el peso de los niños, es continua, y por esto puede tomar infinitos valores.

Cuando los posibles valores numéricos en las clases son infinitos, existe otra manera de definir su tamaño: observando sus valores mínimos y máximos, denominados *límites inferior* y *superior* de la clase, respectivamente. A modo de ejemplo, los límites inferiores de las clases en el ejemplo referido a los pesos de los niños, expresados en kilogramos, son: 38, 40, 42, 44 y 46. Notamos que los límites inferiores van de 2 en 2. Esto corresponde a la amplitud de la clase.

De este modo, en lo que sigue, entenderemos la *amplitud de una clase* como la diferencia entre límites inferiores de clases sucesivas.

Para pensar

¿Cuál podría ser la diferencia entre dos tablas de frecuencia para los mismos datos, donde una utilice muchas más categorías o clases que la otra? ¿Por qué?

Como regla general, podemos decir que, en la medida que aumentamos el número de clases, obtendremos conclusiones más detalladas del comportamiento de la variable de interés. Por otra parte, al disminuir el número de clases, podremos tener una visión más general de este comportamiento. Si bien no es claro el número de clases adecuado en cada caso, en general más de 7 u 8 categorías o clases oscurecen el comportamiento global de la variable, ya que entregan demasiados detalles de su variación que pueden ser irrelevantes. En lo que sigue, discutiremos este punto con mayor profundidad.

2.3.3 Elección de clases en la construcción de tablas de frecuencias para una variable cuantitativa

Como hemos visto, las tablas de frecuencias para datos cuantitativos se construyen de manera similar a las tablas de frecuencias para datos cualitativos, con la diferencia de que el proceso de determinación de clases o categorías es más complejo. Si bien no se espera que alumnos de Educación Básica realicen esta tarea, el futuro docente debe estar preparado para dar lineamientos en la construcción de clases en tablas de frecuencias, y debe conocer las dificultades que los alumnos puedan enfrentar.

Se explicarán dificultades y problemas al construir tablas de frecuencias con datos cuantitativos, en el contexto del siguiente ejemplo:

Considere los siguientes datos que representan el número de visitas a un sitio de Internet, en días sucesivos de un mes.

16	27	26	5	11	33	23	17	15	20	3	14
29	21	23	31	16	8	14	28	19	20	24	35
7	12	22	27	18	20						

Existe una gran variedad de técnicas para determinar la amplitud de las clases que se utilizará y, en general, se basan en la dispersión de los datos. Sin embargo, muchas veces también se considera el contexto y naturaleza de estos. A modo de ejemplo, si bien puede parecer que las clases **0** a **4**, **5** a **9**, **10** a **14**, etc., o **0** a **9**, **10** a **19**, etc. son las más naturales, también es posible utilizar clases como **3** a **7**, **8** a **12**, **13** a **17**, etc. Bajo la primera y la segunda propuesta, se está utilizando el valor mínimo de la variable, número de visitas, como límite inferior de la primera clase. Bajo la tercera propuesta, se está utilizando el mínimo valor de las observaciones.

Una condición que siempre se debe cumplir es que las clases no deben superponerse. Existen, además, ciertos convenios que se recomiendan seguir. Estos son:

1. En la medida de lo posible, todas las clases deben ser de la misma amplitud.

2. En la medida de lo posible, no debe haber clases vacías o con frecuencia **0**. Se debe vigilar que el número de clases vacías corresponda a menos de la mitad del número total de clases.

3. Debe haber suficientes clases, de tal manera que no todos los datos se acumulen en una o dos de ellas.

La **Tabla III.14** muestra tres posibles elecciones de clases, en el ejemplo sobre el número de visitas diarias a cierto sitio de Internet presentado antes.

Clases		
Opción 1	Opción 2	Opción 3
0 a 4	0 a 9	3 a 7
5 a 9	10 a 19	8 a 12
10 a 14	20 a 29	13 a 17
15 a 19	30 a 39	18 a 22
20 a 24		23 a 27
25 a 29		28 a 32
30 a 34		33 a 37
35 a 39		

Tabla III.14: Tres posibles elecciones de clases, en el problema del número de visitas a cierto sitio de Internet.

Estudiaremos, primero, la condición de no superposición de las clases. Esta condición dice que cada valor numérico de los datos debe estar en una y solo una clase. Supongamos que modificamos la Opción 1, como se muestra en la **Tabla III.15**.

Clases	
Opción 1	Opción 1 modificada
0 a 4	0 a 5
5 a 9	5 a 10
10 a 14	10 a 15
15 a 19	15 a 20
20 a 24	20 a 25
25 a 29	25 a 30
30 a 34	30 a 35
35 a 39	35 a 40

Tabla III.15: Definición de clases según Opción 1 y Opción 1 modificada.

Las clases de la Opción 1 modificada tienen ahora límites en común, ¿dónde contaríamos el dato que corresponde a **20** visitas? Podría ser tanto en la cuarta como en la quinta clase, lo cual no debe ocurrir. No debemos, por tanto, utilizar clases que tengan límites en común.

Para estudiar la primera regla recomendada, notamos que en la Opción 1 todos los límites inferiores van de **5** en **5**, luego, todas las clases tienen la misma amplitud. Del mismo modo, corroboramos que todas las clases de las Opciones 2 y 3 tienen la misma amplitud, **10** y **5**, respectivamente. Luego, las **3** propuestas de clases cumplen con la segunda regla.

Para discutir las últimas dos reglas utilizaremos la **Tabla III.16**, relativa al ejemplo que seguimos. Esta tabla contiene cuatro opciones para definir las clases en este problema.

Opción 1		Opción 2		Opción 3		Opción 4	
Clases	Frec.	Clases	Frec.	Clases	Frec.	Clases	Frec.
0 a 4	1	0 a 9	4	3 a 7	3	0 a 19	14
5 a 9	3	10 a 19	10	8 a 12	3	20 a 39	16
10 a 14	4	20 a 29	13	13 a 17	6		
15 a 19	6	30 a 39	3	18 a 22	7		
20 a 24	8			23 a 27	6		
25 a 29	5			28 a 32	3		
30 a 34	2			33 a 37	2		
35 a 39	1						

Tabla III.16: Frecuencias absolutas para cuatro grupos de clases propuestas, en el problema de visitas a cierto sitio de Internet.

Notamos, primero, que ninguna de las clases está vacía, lo que es deseable. Podemos, además, comentar los siguientes puntos: bajo la Opción 1, las clases **0** a **4** y **35** a **39** solo poseen un valor, lo que se considera aceptable. Por otra parte, bajo la Opción 2 se tienen menos categorías, lo que resulta en clases con mayor número de observaciones. Si bien las frecuencias en la Opción 3 difieren de las frecuencias en la Opción 1, el comportamiento de estas es similar. Al tener solo **2** clases, la Opción 4 entrega poca información sobre los datos originales. En efecto, solo podemos leer que alrededor de la mitad de los datos son menores que **20**.

Ejercicios

1. Para cada uno de los siguientes conjuntos de datos, escoja un grupo de clases apropiado para construir una tabla de frecuencias y constrúyala. Justifique su elección de clases.

   a. 5 - 3 - 2 - 1 - 4 - 6 - 9 - 5 - 2 - 3 - 9 - 4 - 6 - 2

   b. 15 - 10 - 24 - 20 - 37- 40 - 16 - 23 - 11 - 24 - 37 - 20

   c. 128 - 113 - 105 - 128 - 97 - 111 - 125 - 101 - 94 - 136

   d. 220 - 340 - 130 - 180 - 230 - 340 - 100 - 112 - 225 - 160 - 175 - 230 - 241 - 100

   e. 5 - 6 - 11 - 0 - 14 - 12 - 8 - 16 - 18 - 9 - 3 - 18 - 0 - 17 - 11 - 3 - 13 - 20

   f. 20 - 10 - 50 - 60 - 40 - 30 - 90 - 70 - 10 - 10 - 60 - 30 - 40 - 20 - 90 - 50

2. Considere la tabla de frecuencias de los sueldos iniciales para 50 profesionales recién graduados:

Sueldo	Frecuencia
menos de $300.000	1
$300.000 a $599.999	16
$600.000 a $899.999	20
$900.000 a $1.199.999	9
$1.200.000 a $1.499.999	4

   a. ¿Cree usted que las clases o categorías han sido definidas de manera adecuada? Explique por qué.
   b. Construya una tabla de frecuencias relativas porcentuales.
   c. ¿Qué porcentaje de profesionales recién graduados recibe entre $300.000 y $899.000?
   d. ¿Qué categoría es la más frecuente? ¿Qué porcentaje de profesionales recién egresados están en ella?

3. Con el objeto de organizar actividades de esparcimiento para los habitantes de una comuna, una municipalidad decidió realizar una encuesta sobre una muestra de 573 habitantes. Los resultados reportados se muestran en la siguiente tabla.

Edad	Frecuencia absoluta
0 a 9 años	73
10 a 19 años	74
20 a 39 años	178
40 a 49 años	94
50 a 59 años	70
60 o más años	84

   a. Note que no todas las categorías o clases tienen la misma amplitud, lo que, en ciertos casos, puede ser razonable. Considerando el interés de la municipalidad, indique a qué puede deberse la elección de clases utilizada.
   b. ¿Qué otra elección de clases puede proponer? ¿Puede construir la tabla con las clases propuestas a partir de la tabla reportada en el estudio?

4. Busque un conjunto de datos cuantitativos en Internet y discuta posibles definiciones de clases para ellos.

2.4. Tablas de frecuencias para dos variables cualitativas

2.4.1 Construcción de tablas de frecuencias para dos variables cualitativas

Retomemos nuevamente el problema de determinar los deportes favoritos de los niños y supongamos que, además de conocer los deportes favoritos en el curso, quisiéramos saber qué ocurre cuando consideramos separadamente niños y niñas. ¿Serán, en general, diferentes los deportes preferidos por niños y niñas? En este caso, al registrar las respuestas también deberemos anotar el sexo del niño, que en ese caso puede obtenerse a partir de la **Tabla III.1**.

Para organizar la información recolectada, podemos construir diferentes tablas. Una posibilidad es construir una tabla en base a lo que ya conocemos para una sola variable. Es decir, llamamos categoría a cada combinación de deporte favorito y sexo del niño. De este modo, hemos creado $4 \times 2 = 8$ categorías, que corresponden a todas las posibles combinaciones de 4 deportes y 2 sexos. Bajo este enfoque, podríamos reportar una tabla como la que se muestra a continuación.

Deporte favorito y sexo	Frecuencia absoluta
Fútbol, niño	4
Fútbol, niña	2
Básquetbol, niño	4
Básquetbol, niña	0
Gimnasia rítmica, niño	1
Gimnasia rítmica, niña	2
Tenis, niño	1
Tenis, niña	1

Tabla III.17: Frecuencias de cada categoría formada por una combinación de deporte favorito y sexo del niño.

La **Tabla III.17** nos permite responder preguntas como ¿cuántas niñas hay en el curso cuyo deporte favorito es el tenis?, y su respuesta se lee directamente de la fila 8 de la tabla.

Consideremos ahora las preguntas:

- ¿Cuántas niñas hay en el curso?
- ¿Qué diferencia hay entre el número de niñas a las que les gusta el fútbol y el número de niñas a las que les gusta el tenis?

Para responder a la primera pregunta, es posible sumar las frecuencias fila por medio a partir de la segunda, es decir, sumar $2 + 0 + 2 + 1 = 5$ niñas. Para responder la segunda pregunta, debemos restar dos filas alejadas, la segunda y la octava, obteniendo que la diferencia a favor del fútbol es $2 - 1 = 1$ niña. Como vemos, si bien la tabla contiene los datos para responder a las preguntas, esta no muestra a simple vista la información pedida.

Para obtener este tipo de información, resulta más útil entregar una *tabla de frecuencias de dos entradas*, como la que se muestra en la **Tabla III.18**. En ella, reconocemos la presencia de dos variables: deporte favorito y sexo del niño. Las filas corresponden a las categorías de la variable deporte favorito, mientras que las columnas indican las categorías asociadas al sexo.

	Sexo		
Deporte favorito	Mujer	Hombre	Total
Fútbol	2	4	6
Básquetbol	0	4	4
Gimnasia rítmica	2	1	3
Tenis	1	1	2
Total	5	10	15

Tabla III.18: Tabla de dos entradas de las frecuencias absolutas de deporte favorito y sexo de los niños del curso.

Cada celda interior de la tabla contiene el número de observaciones que pertenece simultáneamente a las categorías indicadas por su fila y su columna. De las celdas interiores leemos, por ejemplo, que hay **2** niñas cuyo deporte favorito es el fútbol. Si retomamos la pregunta ¿cuántas niñas hay en el curso?, en esta tabla basta con mirar la columna Mujer y sumar todos los deportes. La operación aditiva es la misma que se hizo anteriormente a partir de la **Tabla III.17**, $2 + 0 + 2 + 1 = 5$ niñas, sin embargo, la información necesaria está ahora mucho más a mano. Más aun, esta suma suele ponerse en la última fila de la tabla, como se ha hecho en la **Tabla III.18**.

Del mismo modo, si consideramos la pregunta ¿qué diferencia hay entre el número de niñas a las que le gusta el fútbol y a las que les gusta el tenis?, podemos responder mirando únicamente la columna titulada Mujer.

Tal como ocurrió al determinar el número de niñas en el curso, notamos que a partir de la tabla de frecuencias absolutas de dos entradas es posible recuperar las frecuencias absolutas contenidas en las tablas de frecuencias individuales de cada variable. Si consideramos, por ejemplo, la variable deporte favorito, es posible obtener la frecuencia absoluta de cada deporte sumando el número de niños y de niñas por fila. Por ejemplo, para fútbol, $2 + 4 = 6$ niños del curso lo prefieren. Repitiendo esto para cada deporte, podríamos construir la tabla de frecuencias absolutas para deporte favorito, con los valores de la última columna de la **Tabla III.18**.

Sin embargo, una tabla de frecuencias de dos entradas, como la **Tabla III.18**, nos entrega información adicional a la que entregan las tablas de frecuencia de una sola variable. Por ejemplo, considere nuevamente la **Tabla III.5**, presentada anteriormente, que entrega solo las frecuencias sobre deportes, y en la cual podemos leer que un total de **6** hombres prefieren el fútbol. Adicionalmente, considere la **Tabla III.19** siguiente, que muestra solo las frecuencias del sexo de los niños, y en la cual podemos leer que, del total de alumnos, **5** son mujeres.

Sexo	Número de niños
Hombres	10
Mujeres	5
Total	15

Tabla III.19: Tabla de frecuencias del sexo de los niños.

Para pensar

¿Se puede determinar a partir de las **Tablas III.5** y **III.19** el número de niñas que prefiere el fútbol?

Preguntas interesantes, en este caso, pueden ser ¿a quiénes les gustará más el tenis, a las niñas o a los niños? ¿tendrán las niñas y los niños los mismos deportes favoritos? En la **Tabla III.18** vemos que, para cada deporte, los números de niños y niñas que los prefieren son diferentes. Con respecto a la primera pregunta, en la tabla vemos que hay igual número de niños y niñas a los que les gusta el tenis. Sin embargo, no podemos sacar conclusiones basadas únicamente en este hecho.

Para responder las preguntas, debemos calcular frecuencias relativas (o relativas porcentuales), por sexo. Con respecto a la primera pregunta que nos hemos hecho ¿a quiénes les gustará más el tenis, a las niñas o a los niños?, de la **Tabla III.18** podemos calcular que un:

$$100 \bullet \frac{1}{5}\% = 20\%,$$

de las niñas prefiere el tenis, mientras que solo un:

$$100 \bullet \frac{1}{10}\% = 10\%,$$

de los niños prefiere este mismo deporte. Podemos concluir que, aunque el número de niños que prefiere el tenis es igual al de niñas, en realidad, un mayor porcentaje de ellas lo prefiere, lo que muestra una mayor preferencia de las niñas por el tenis. Al referirnos anteriormente al fútbol, también habíamos notado que, aunque el número de niñas que prefiere el fútbol es la mitad del número de niños, en realidad, este deporte es igualmente favorito entre niñas y niños, puesto que ambas frecuencias relativas porcentuales son iguales a **40%**.

La segunda pregunta, ¿tendrán las niñas y los niños las mismas preferencias?, requiere determinar si los porcentajes de niñas y de niños que prefieren cada deporte son similares. Podemos, entonces, repetir los cálculos realizados para el tenis en cada categoría, dividiendo las frecuencias absolutas de la tabla por el número total de niños o de niñas, según corresponda. Las frecuencias relativas porcentuales se muestran en la **Tabla III.20**. Las frecuencias relativas porcentuales de cada columna deben siempre sumar **100%**.

	Sexo	
Deporte favorito	Mujer	Hombre
Fútbol	40%	40%
Básquetbol	0%	40%
Gimnasia rítmica	40%	10%
Tenis	20%	10%
Total	100%	100%

Tabla III.20: Tabla de frecuencias relativas porcentuales de cada deporte por cada sexo.

En la tabla, vemos que niños y niñas tienen distintos deportes favoritos. Podemos decir, por ejemplo, que el porcentaje de niñas para las que su deporte favorito es la gimnasia rítmica es bastante mayor que el porcentaje de niños que prefiere este mismo deporte, siendo estos 40% y 10%, respectivamente.

De este modo, una tabla de frecuencias de dos entradas nos entrega información sobre la relación que existe entre dos variables; en el ejemplo, la relación que existe entre el deporte favorito y el sexo de los niños del curso. Es importante notar que también es posible plantear la pregunta de forma inversa: ¿difieren los porcentajes de niños y niñas en cada deporte? En este caso, debemos calcular las frecuencias relativas para cada deporte, en lugar de para cada sexo. A modo de ejemplo, dentro de los niños y niñas a los que les gusta el fútbol, en la **Tabla III.18** vemos que el porcentaje de niños es:

$$100 \cdot \frac{4}{6}\%$$

o aproximadamente un 66,7%, y el de niñas es:

$$100 \cdot \frac{2}{6}\%$$

o aproximadamente un 33,3%.

Podemos repetir este cálculo para todos los deportes y obtener la **Tabla III.21**. En ella, notemos que los porcentajes suman 100% por fila.

	Sexo		
Deporte favorito	Mujer	Hombre	Total
Fútbol	33,3%	66,7%	100%
Básquetbol	0%	100%	100%
Gimnasia rítmica	66,7%	33,3%	100%
Tenis	50%	50%	100%

Tabla III.21: Tabla de frecuencias porcentuales de cada sexo, por deporte favorito.

2.4.2 Lectura e interpretación de tablas de frecuencias para dos variables

Es habitual encontrar tablas de frecuencias de dos entradas cuando se desea mostrar la relación existente entre dos variables. En particular, la **Tabla III.22** corresponde a una tabla de frecuencias relativas porcentuales tomada del informe de la "5ta encuesta nacional de la juventud". La tabla se refiere a la relación entre la jornada de trabajo preferida y el sexo de los jóvenes.

		Sexo	
Tipo de jornada	Total	Hombre	Mujer
Jornada completa	60,7%	67,0%	52,3%
Media jornada	18,2%	12,9%	25,3%
Part-time/por horas	16,6%	15,3%	18,3%
Otro tipo de jornada	3,5%	3,8%	3,0%
No responde	1,0%	0,9%	1,2%

Tabla III.22: Tabla de frecuencias porcentuales de cada tipo de jornada preferida, por sexo.

¿Cómo se interpreta el valor **67%** que leemos en la tercera columna de la **Tabla III.22**? Para entender esta tabla, primero debemos determinar el tipo de porcentajes que se muestran: los porcentajes de preferencias por cada tipo de jornada, de hombres y mujeres por separado, o los porcentajes de hombres y mujeres, para cada tipo de jornada por separado. Esto se realiza obteniendo las sumas de los porcentajes mostrados.

En la **Tabla III.22**, los porcentajes suman **100%** por columna [8], lo que nos dice que la tabla muestra los porcentajes de preferencias por cada tipo de jornada, de hombres y mujeres por separado. Leemos, por ejemplo, que **67%** de los hombres prefiere un trabajo de jornada completa, bajando este porcentaje a **52,3%** para las mujeres. Notamos que, en cambio, las mujeres prefieren, en mayor porcentaje que los hombres, un trabajo de media jornada, **25,3%** versus **12,9%**. Vemos así la diferencia entre las preferencias de hombres y mujeres.

Se debe considerar que, a partir de una tabla como la **Tabla III.22**, que muestra porcentajes por columna, no es posible recuperar los porcentajes por fila, esto es, determinar, por ejemplo, qué porcentaje de los jóvenes que prefieren jornada completa corresponden a hombres y qué porcentaje, a mujeres. La decisión sobre qué tipo de porcentajes mostrar en una tabla, por fila o por columna, dependerá de cada estudio en particular. En este caso, lo que se desea es comparar las preferencias de hombres y mujeres, lo que determina que se deben comparar los porcentajes de hombres y mujeres por separado.

Para pensar

Observando la **Tabla III.22**, en su totalidad, ¿cree usted que existe alguna asociación entre el tipo de jornada preferida y el sexo del entrevistado?

[8] Las sumas son aproximadas, debido al redondeo.

Examinaremos, a continuación, otro ejemplo, tomado esta vez, del informe "Implementación curricular en el aula. Matemáticas. Primer ciclo básico (NB1 y NB2)"[9], preparado por el Ministerio de Educación, que se basa en observaciones recogidas en los años 2001 y 2002.

La **Tabla III.23** resume la distribución del porcentaje de tiempo destinado a cada uno de los bloques de contenido en primero y segundo básico, dentro del total de tiempo dedicado a la asignatura de Matemática en un año escolar. Esta tabla fue construida en base a información recolectada desde libros de clase y cuadernos.

Bloque de contenidos	1º básico	2º básico
Apresto	13,1%	–
Números naturales	43,5%	31,7%
Operaciones aritméticas	25,2%	30,3%
Geometría	3,3%	5,9%
Orientación en el espacio	–	0,4%
Fracciones	0,8%	1,9%
Resolución de problemas	3,3%	8,4%
Evaluaciones	6,6%	9,6%
Otros temas	4,2%	4,9%
Sin registro	–	6,9%

Tabla III.23: Porcentaje de tiempo destinado a cada uno de los bloques de contenido en primero y segundo básico, dentro del total del tiempo dedicado a la asignatura de Matemática en un año escolar.

Nuevamente, los porcentajes suman **100%** por columna, por lo que, a modo de ejemplo, leemos que en primero básico, el **43,5%** del tiempo fue dedicado a estudiar los números naturales, bajando a **31,7%** en segundo básico. Notamos, en cambio, que en segundo básico se dedica un mayor porcentaje del tiempo a estudiar operaciones aritméticas y resolución de problemas que en primero básico; **30,3%** versus **25,2%** en operaciones aritméticas, y **8,4%** versus **3,3%** en resolución de problemas. Vemos así la evolución del énfasis dado a cada contenido cubierto de un nivel escolar a otro.

[9] Tomado de la página web de Educar Chile. http://www.educarchile.cl/Userfiles/P0001/File/Estudio%20curricularMatemat.pdf

2.4.3 Dificultades en la construcción e interpretación de tablas de frecuencias de dos entradas: frecuencias relativas porcentuales

Una de las dificultades en el manejo de tablas de frecuencias de dos entradas se relaciona al cálculo e interpretación de frecuencias relativas, al momento de determinar si estas corresponden a frecuencias relativas calculadas por fila, por columna o totales.

Consideremos la información que se muestra en la **Tabla III.24**, que corresponde a las personas a bordo del Titanic al momento de su hundimiento, clasificadas de acuerdo a la clase en la que viajaban y su sobrevivencia.

	Clase				
Sobrevivencia	Primera	Segunda	Tercera	Tripulación	Total
Sobrevivientes	203	118	178	212	711
No sobrevivientes	122	167	528	673	1.490
Total	325	285	706	885	2.201

Tabla III.24: Tabla de frecuencias absolutas de los pasajeros del Titanic, según la clase en que viajaron y su sobrevivencia.

Las siguientes frecuencias relativas porcentuales suenan similares, pero son diferentes:

- El porcentaje de pasajeros que viajaban en primera clase y que sobrevivieron. Este se calcula como $\frac{203}{2.201}$ o aproximadamente, **9,2%**.
- El porcentaje de los pasajeros de primera clase que sobrevivieron. Este se calcula como $\frac{203}{325}$ o aproximadamente, **62,5%**.
- El porcentaje de sobrevivientes que eran de primera clase. Este sería $\frac{203}{711}$ o aproximadamente, **28,6%**.

En cada caso, se debe identificar correctamente el grupo de pasajeros dentro del cual se desea obtener el porcentaje. Es usual que el grupo al que nos referimos corresponda a un subgrupo de la muestra. La redacción utilizada juega un rol fundamental para entender estas situaciones.

En el primer porcentaje calculado, el grupo considerado es el grupo de todas las personas a bordo del Titanic. En el segundo porcentaje calculado, el grupo considerado es el subgrupo de pasajeros de primera clase. Es posible identificar esta situación a partir de la palabra "de". Dado que dice "de primera clase", estamos obligados a considerar solo esta clase, por lo que el denominador debe ser el total de pasajeros en ella, **325**. En el último porcentaje calculado, el subgrupo considerado corresponde al grupo de sobrevivientes a la tragedia. Dado que dice "de sobrevivientes", estamos obligados a considerar solo esa categoría, por lo que el denominador debe ser el total de las personas que sobrevivieron, **711**.

Este punto fue tocado en la discusión de las **Tablas III.22** y **III.23**, sobre la preferencia de jornada laboral y el tiempo dedicado a contenidos escolares, respectivamente, donde debimos deducir el subgrupo sobre el cual se calcularon los porcentajes reportados. En la **Tabla III.22**, notamos que los subgrupos sobre los cuales se obtuvieron los porcentajes corresponden a hombres y mujeres. En la **Tabla III.23**, notamos que los subgrupos sobre los cuales se obtuvieron los porcentajes reportados corresponden a cada nivel escolar, primero y segundo básico.

¿Cómo elegir el subgrupo sobre el cual obtener las frecuencias relativas? Consideremos nuevamente el ejemplo en la **Tabla III.24** sobre los pasajeros del Titanic. Una tarea interesante corresponde a explorar si existe una relación entre la sobrevivencia de los pasajeros y la clase en la que viajaban. Tenemos dos posibilidades para explorar esta relación. La primera corresponde a observar la distribución de la clase en cada condición de sobrevivencia. Esto equivale a calcular las frecuencias relativas por cada fila, donde los subgrupos se definen por la condición de sobrevivencia. Las sumas de los porcentajes en cada fila deben ser iguales a **100%**. Estas frecuencias relativas se muestran en la **Tabla III.25**.

A modo de ejemplo, el porcentaje de **8,2%** en la segunda columna, última fila, corresponde al porcentaje de los no sobrevivientes que viajaban en primera clase. Esta interpretación surge de la observación de que los porcentajes suman **100%** en cada fila, por lo que los subgrupos considerados son los pasajeros sobrevivientes, en la primera fila, y los no sobrevivientes, en la segunda.

	Clase				
Sobrevivencia	Primera	Segunda	Tercera	Tripulación	Total
Sobrevivientes	28,6%	16,6%	25,0%	29,8%	100%
No sobrevivientes	8,2%	11,2%	35,4%	45,2%	100%

Tabla III.25: Tabla de frecuencias relativas porcentuales de los pasajeros del Titanic. Las frecuencias relativas han sido calculadas para cada condición de sobrevivencia y suman 100% por fila.

Una segunda alternativa para estudiar la presencia de una asociación entre la clase de viaje y la sobrevivencia corresponde a explorar la distribución de la condición de sobrevivencia dentro de cada clase, como se muestra en la **Tabla III.26**. En este caso, calculamos las frecuencias relativas porcentuales por columna. Esta información nos muestra si la sobrevivencia es aproximadamente la misma en cada una de las cuatro clases. Notemos que ahora los porcentajes suman **100%** por columna.

	Clase			
Sobrevivencia	Primera	Segunda	Tercera	Tripulación
Sobrevivientes	62,5%	41,4%	25,2%	24,0%
No sobrevivientes	37,5%	58,6%	74,8%	76,0%
Total	100,0%	100,0%	100,0%	100,0%

Tabla III.26: Tabla de frecuencias relativas de los pasajeros del Titanic. Las frecuencias relativas han sido calculadas según clase.

¿Cuál de las dos tablas de frecuencias relativas es más apropiada para explorar la posible asociación entre las variables de sobrevivencia y clase? Sin considerar el contexto de las variables, ambas tablas, **III.25** y **III.26**, pueden ser utilizadas para la exploración de la asociación. Sin embargo, se debe tener mucho cuidado.

En efecto, en la **Tabla III.25** puede leerse equivocadamente que la tripulación tenía más posibilidades de sobrevivir, dado que dentro de las clases que lo hicieron, la clase tripulación es la de mayor porcentaje (**29,8%** versus **28,6%**, **16,6%** y **25%**). Sin embargo, en la **Tabla III.24** notamos que el número de pasajeros de la tripulación es el más grande del barco, **885** personas, y por eso

el porcentaje anterior es más alto. De hecho, de la **Tabla III.25** también podemos leer que entre los no sobrevivientes, el mayor porcentaje también corresponde a la tripulación (**45,2%** versus **8,2%**, **11,2%** y **35,4%**).

Por otra parte, si miramos la **Tabla III.26**, notamos que si un pasajero es de primera clase, sus posibilidades de sobrevivir eran de **62,5%**, mayores que las posibilidades de sobrevivir en las clases restantes (**62,5%** versus **41,4%**, **25,2%** y **24%**). De la **Tabla III.24**, vemos que esto no se debe a que haya más pasajeros sobrevivientes (**711** sobrevivientes versus **1.490** no sobrevivientes).

En nuestro ejemplo, la pregunta de interés puede ser formulada como ¿depende la sobrevivencia de un pasajero de la clase en la que viajaba? Si el problema está formulado de esta forma, se está pensando en que la clase podría explicar la sobrevivencia de los pasajeros. En este caso, debiésemos comparar los porcentajes de sobrevivencia en cada una de las cuatro clases: primera, segunda, tercera y tripulación. Es decir, los subgrupos sobre los que se calcularán los porcentajes corresponden a las diferentes clases. Esta información está contenida en la **Tabla III.26**. Observamos en la tabla que el porcentaje de sobrevivientes va disminuyendo en la medida que avanzamos desde la primera clase a la tripulación. Esto sugiere una asociación entre la sobrevivencia y la clase en que viajaba el pasajero.

Lo anterior puede ser formalizado diciendo que existen principalmente dos tipos de tablas de frecuencias de dos entradas: *conjuntas* y *condicionales*. Para entenderlas, repetiremos el tipo de análisis realizado a los pasajeros del Titanic, pero esta vez consideraremos los datos en el informe de la "Encuesta nacional de hábitos de actividad física y deportes en la población chilena de 18 años y más" [10] de 2012, referentes al interés de los entrevistados en la actividad física y el deporte, y sus hábitos de práctica. La **Tabla III.27** muestra el número de entrevistados según su interés y hábitos, y su sexo.

	Sexo		
Interés y práctica	Masculino	Femenino	Total
Le interesa el deporte y la actividad física, y los practica.	778	859	1.637
No le interesa el deporte y la actividad física, pero los practica por obligación.	50	44	94
Le interesa el deporte y la actividad física, pero no los practica.	946	2.084	3.030
No le interesa el deporte y la actividad física, y no los practica.	316	693	1.009
No sabe o no responde.	2	7	9
Total	2.092	3.687	5.779

Tabla III.27: Frecuencias absolutas de entrevistados según interés en la actividad física y el deporte, su práctica, y el sexo de los entrevistados.

[10] Información tomada de la página web del Instituto Nacional de Deportes. http://www.ind.cl/estudios-e-investigacion/investigaciones/Documents/2012/Encuesta%20Act%20Fisica/encuesta-act-fisica-2012.pdf

Repitiendo el análisis que realizamos en el ejemplo sobre el Titanic, si deseamos conocer el porcentaje de entrevistados a los que les interesa el deporte y la actividad física y los practican, y que son hombres, este corresponde a $\frac{778}{5.779}$ x 100%; aproximadamente un 13,5% de los entrevistados. Esta frecuencia relativa porcentual se denomina *frecuencia relativa porcentual conjunta* y una manera de reconocerla es notar que se ha utilizado la palabra "y" para unir dos categorías, una asociada al interés y práctica de deporte del entrevistado, y otra asociada a su sexo.

Podemos repetir este cálculo para todas las celdas en la **Tabla III.27** y obtendremos la **Tabla III.28**. Esta tabla se denomina *tabla de frecuencias relativas porcentuales conjunta*, puesto que muestra los porcentajes del interés por el deporte y la actividad física y su práctica, y el sexo, a la vez. Una manera de reconocer que la **Tabla III.28** es una tabla conjunta corresponde a notar que la suma total de porcentajes, que se muestra en la celda de las últimas fila y columna, corresponde a 100%.

	Sexo		
Interés y práctica	Masculino	Femenino	Total
Le interesa el deporte y la actividad física, y los practica.	13,5%	14,9%	28,3%[11]
No le interesa el deporte y la actividad física, pero los practica por obligación.	0,9%	0,8%	1,7%
Le interesa el deporte y la actividad física, pero no los practica.	16,4%	36,0%	52,4%
No le interesa el deporte y la actividad física, y no los practica.	5,5%	12,0%	17,5%
No sabe o no responde.	0,0%	0,1%	0,1%
Total	36,2%[12]	63,8%	100%

Tabla III.28: Tabla de frecuencias relativas porcentuales conjunta del interés en la actividad física y el deporte, y su práctica, y el sexo de los entrevistados.

De la **Tabla III.28** leemos también, por ejemplo, que aproximadamente un 36,0% de los entrevistados corresponden a personas a las que les interesa el deporte y la actividad física, pero no los practican, y son mujeres. También leemos que a aproximadamente un 28,3% de las personas entrevistadas les interesa la actividad física y el deporte, y los practican, lo que se lee en la columna de la derecha, que suma las frecuencias relativas porcentuales de hombres y mujeres para cada interés y práctica. Notamos que esta última frecuencia relativa porcentual no corresponde a una frecuencia relativa conjunta, dado que involucra solo una variable: el interés por la actividad física y el deporte y su práctica. Es por esto que esta probabilidad se indica en la columna titulada "Total", y no en el cuerpo de la tabla de frecuencias relativas conjuntas.

[11] La suma de la fila que se muestra no es exacta debido al redondeo.
[12] La suma de la columna que se muestra no es exacta debido al redondeo.

Por otra parte, volviendo a la **Tabla III.27**, notamos que el porcentaje de hombres a quienes les interesa el deporte y la actividad física, y los practica corresponde a $\frac{778}{2.092}$ x 100%, o aproximadamente un 37,2%. Notamos que estamos dividiendo el número de observaciones en una celda por el total de observaciones en su columna, dado que el subgrupo a considerar es el grupo de hombres. Del mismo modo, podemos proceder con las celdas restantes, obteniendo la **Tabla III.29**. Esta tabla corresponde a una *tabla de frecuencias relativas porcentuales condicionales por columna*, ya que se divide por el total de la columna. También podemos decir que la tabla muestra los porcentajes por columna. Una manera de reconocer que estamos frente a una tabla condicional por columna es notando que la suma de los porcentajes en cada columna es 100%.

	Sexo	
Interés y práctica	Masculino	Femenino
Le interesa el deporte y la actividad física, y los practica.	37,2%	23,3%
No le interesa el deporte y la actividad física, pero los practica por obligación.	2,4%	1,2%
Le interesa el deporte y la actividad física, pero no los practica.	45,2%	56,5%
No le interesa el deporte y la actividad física, y no los practica.	15,1%	18,8%
No sabe o no responde.	0,1%	0,2%
Total	100%	100%

Tabla III.29: Tabla de frecuencias relativas porcentuales condicional del interés en la actividad física y el deporte y su práctica, para cada sexo de los entrevistados.

De la **Tabla III.29** podemos leer, por ejemplo, que aproximadamente al **56,5%** de las mujeres les interesa el deporte y la actividad física, pero no los practica.

Tal como hemos construido una tabla de frecuencias relativas porcentuales condicionales por columna, podemos construir una tabla de frecuencias relativas porcentuales por filas. A partir de la **Tabla III.27**, podemos obtener, por ejemplo, que dentro de los entrevistados a los que les interesa el deporte y la actividad física y los practica, el porcentaje de hombres es $\frac{778}{1.637}$ x 100%, o aproximadamente un 47,5%. Ahora estamos dividiendo el número de observaciones en una celda por el total de observaciones en su fila. Del mismo modo, podemos proceder con las celdas restantes, obteniendo la **Tabla III.30**. Esta tabla corresponde a una *tabla de frecuencias relativas porcentuales condicionales por fila*, dado que se divide por el total de la fila. También podemos decir que la tabla muestra los porcentajes por fila. Una manera de reconocer que estamos frente a una tabla condicional por fila es notando que la suma de los porcentajes en cada fila es 100%[13].

[13] Los porcentajes se encuentran redondeados al primer decimal, lo que puede incidir en sus sumas.

	Sexo		
Interés y práctica	Masculino	Femenino	Total
Le interesa el deporte y la actividad física, y los practica.	47,5%	52,5%	100%
No le interesa el deporte y la actividad física, pero los practica por obligación.	53,2%	46,8%	100%
Le interesa el deporte y la actividad física, pero no los practica.	31,2%	68,8%	100%
No le interesa el deporte y la actividad física, y no los practica.	31,3%	68,7%	100%
No sabe o no responde.	22,2%	77,8%	100%

Tabla III.30: Tabla de frecuencias relativas porcentuales condicional del sexo de los entrevistados, para cada tipo de interés en el deporte y actividad física y su práctica.

De la **Tabla III.30** leemos, por ejemplo, que entre las personas a quienes no les interesa el deporte y la actividad física, pero los practica por obligación, un **46,8%** son mujeres.

Al igual como lo discutimos con la **Tabla III.22**, se debe tener presente que, en general, no es posible construir una tabla de frecuencias relativas por fila únicamente a partir de la correspondiente tabla de frecuencias relativas por columna, y viceversa. A modo de ejemplo, a partir de la **Tabla III.29** no podríamos determinar el porcentaje de entrevistados a los que les interesa el deporte y la actividad física y los practican, que son mujeres, así como tampoco podríamos conocer, a partir de la **Tabla III.30**, el porcentaje de las mujeres a las que les interesa el deporte y la actividad física y los practican.

Según las denominaciones que hemos utilizado, podemos afirmar que las **Tablas III.25** y **III.26**, referentes al hundimiento del Titanic, corresponden a tablas condicionales por fila y por columna, respectivamente.

En resumen

- Una *tabla de frecuencias de dos entradas* corresponde a una tabla de frecuencias para dos variables, donde, las filas corresponden a una de ellas y las columnas, a la otra.
- Cada categoría en la tabla corresponde a una combinación de las categorías de las dos variables involucradas.
- En una tabla de frecuencias de dos entradas, lo más común es presentar frecuencias relativas (o relativas porcentuales), las que pueden ser calculadas por fila o columna (*frecuencias relativas condicionales*) o sobre el total de las observaciones (*frecuencias relativas conjuntas*), según lo que se desee comunicar.
- Es deseable contar con las frecuencias absolutas para poder extraer conclusiones correctas.

Ejercicios

1. Los alumnos de un curso universitario fueron entrevistados acerca de sus tendencias políticas descritas como "Izquierda", "Centro" o "Derecha". La información puede ser resumida en la siguiente tabla:

	Tendencia política			
Sexo	Izquierda	Centro	Derecha	Total
Hombre	35	36	6	77
Mujer	50	44	21	115
Total	85	80	27	192

   a. ¿Qué porcentaje del curso es hombre?
   b. ¿Qué porcentaje del curso se considera de "Centro"?
   c. ¿Qué porcentaje de los hombres del curso se considera de "Derecha"?
   d. ¿Qué porcentaje de todos los alumnos del curso son hombres que se consideran de "Derecha"?

2. Considere nuevamente la tabla del ejercicio anterior sobre las tendencias políticas de alumnos universitarios.

   a. Encuentre los porcentajes de las tendencias políticas entre las mujeres.
   b. Encuentre los porcentajes de las tendencias políticas entre los hombres.
   c. ¿Existe alguna razón para creer que política y sexo están relacionados? Explique.

3. Este problema estudia la precisión de los pronósticos del tiempo reportados por los medios de comunicación. La tabla que se presenta compara los pronósticos climáticos con las condiciones climáticas observadas para el mismo período del pronóstico en una ciudad, durante un año:

	Clima observado	
Pronóstico	Lluvia	No lluvia
Lluvia	27	63
No lluvia	7	268

   a. ¿Qué porcentaje de las veces el pronóstico estuvo en lo correcto?
   b. ¿Qué porcentaje de los días que llovió, el pronóstico estuvo en lo correcto?
   c. ¿Qué porcentaje de los días que no llovió, el pronóstico estuvo en lo correcto?
   d. ¿Se percibe evidencia de que existe una relación entre las condiciones climáticas y la precisión de los pronósticos?

4. Consideremos la siguiente tabla, obtenida a partir de los resultados de la "5ta encuesta nacional de la juventud" de 2006. La tabla muestra los factores que los jóvenes indican que los llevan a no estar estudiando.

Razón principal	2003	2006
Terminó educación	18,3%	32,0%
Problemas económicos	27,3%	22,3%
Decidí trabajar	22,6%	16,3%
Por cuidar a mi hijo	14,3%	11,0%
Dificultades académicas	3,0%	3,4%
No me interesó estudiar	2,9%	2,7%
En preuniversitario	2,6%	1,3%
Tuve que ayudar en casa	1,3%	1,1%
Problemas de conducta	0,4%	0,9%
Por enfermedad	1,7%	0,3%
Otra razón	5,8%	5,2%
No responde	0,0%	3,6%

a. ¿Qué porcentaje de jóvenes dejó de estudiar por problemas económicos en el año 2003? ¿y en el año 2006?

b. ¿Qué porcentaje de jóvenes dejó de estudiar porque decidió trabajar en el año 2003? ¿y en el año 2006?

c. Extraiga dos conclusiones comparativas generales de las razones mencionadas entre los años 2003 y 2006.

5. La siguiente tabla de frecuencias relativas porcentuales fue construida en base a datos de la "Encuesta nacional de hábitos de actividad física y deportes en la población chilena de 18 años y más"[14] del año 2012, y se refiere a la edad y hábitos deportivos.

Edad	18 a 25	26 a 35	36 a 45	46 a 55	56 a 65	Más de 65
No practicante	54,0%	63,7%	73,3%	74,0%	82,4%	78,3%
Practicante	46,0%	36,3%	26,7%	26,0%	17,6%	21,7%

a. ¿Qué representa el valor **54%** en la primera celda de la tabla? ¿y el valor **82,4%** en la primera fila y quinta columna?

b. ¿Qué porcentaje de entrevistados entre **36** y **45** años practican deporte?

c. ¿Puede conocer, a partir de la tabla, el porcentaje de entrevistados que practican deporte que tienen entre **26** y **35** años?

d. ¿Puede conocer, a partir de la tabla, el porcentaje de entrevistados entre **18** a **35** años que no practican deporte?

[14] Información tomada de la página web del Instituto Nacional de Deportes. http://www.ind.cl/estudios-e-investigacion/investigaciones/Documents/2012/Encuesta%20Act%20Fisica/encuesta-act-fisica-2012.pdf

2.4.4 La paradoja de Simpson

Hemos visto que las tablas de frecuencias relativas de dos entradas nos hablan de la relación entre dos variables. La importancia de usar este tipo de tablas es que, en ocasiones, una tabla de frecuencias individual para una de las variables, sin considerar el desglose según la otra, puede oscurecer las conclusiones correctas.

Esto también puede ser cierto al estudiar la relación entre dos variables. En efecto, es posible que exista una tercera variable que, al no ser tomada en cuenta, oscurezca o confunda la relación entre las dos primeras. A modo de ejemplo, consideremos la siguiente situación:

Dos pilotos, Marina y Julio, discuten sobre sus habilidades para dirigir aviones basándose en la información sobre sus vuelos contenida en la **Tabla III.31**. La tabla entrega los porcentajes de vuelos que cada uno ha realizado a tiempo y con retraso, sus últimos **120** vuelos. Se han obtenido las frecuencias relativas porcentuales para cada uno de los pilotos por separado, es decir, se ha dividido por los totales de fila:

	Llegada		
Piloto	A tiempo	Con retraso	Total
Marina	83%	17%	100%
Julio	78%	22%	100%

Tabla III.31: Frecuencias relativas porcentuales de llegadas (a tiempo o con retraso) de Marina y Julio.

De acuerdo a la tabla, Marina argumenta que ella es mejor piloto que Julio, ya que pudo aterrizar a tiempo el **83%** de sus últimos **120** vuelos, comparado con solo el **78%** de Julio.

Julio, sin embargo, argumenta que la información utilizada distorsiona la realidad dado que sus puntualidades dependen de si el vuelo es realizado de día o de noche. Por esto, decide construir dos tablas por separado, una para los vuelos de día y otra para los vuelos de noche, como se muestra en las **Tablas III.32** y **III.33.** Julio ahora argumenta que él es mejor piloto que Marina, pues sus vuelos han llegado a tiempo un mayor porcentaje de veces, tanto de día (**95%** de Julio versus **90%** de Marina), como de noche (**75%** de Julio versus **50%** de Marina).

	Llegada de vuelos de día		
Piloto	A tiempo	Con retraso	Total
Marina	90%	10%	100%
Julio	95%	5%	100%

Tabla III.32: Frecuencias relativas porcentuales de llegadas (a tiempo o con retraso) de vuelos de día, de Marina y Julio.

	Llegada de vuelos de noche		
Piloto	A tiempo	Con retraso	Total
Marina	50%	50%	100%
Julio	75%	25%	100%

Tabla III.33: Frecuencias relativas porcentuales de llegadas (a tiempo o con retraso) de vuelos de noche, de Marina y Julio.

Para pensar

¿Cómo puede explicarse esta aparente contradicción entre lo que señala Marina y lo que señala Julio?

Al comparar las **Tablas III.32** y **III.33**, vemos que existe una relación entre los desempeños de los pilotos y las condiciones en que se realiza el vuelo: día o noche. En efecto, ambos pilotos llegan a tiempo un mayor porcentaje de veces al volar de día que de noche. Esta relación debe ser considerada, que es lo que se ha hecho en las **Tablas III.32** y **III.33**, para llegar a las conclusiones correctas.

La **Tabla III.34** ayuda a entender el origen del malentendido. La tabla muestra el número de vuelos (frecuencias absolutas) de día y de noche realizados por cada piloto.

	Llegada		
Piloto	Día	Noche	Total
Marina	100	20	**120**
Julio	20	100	**120**

Tabla III.34: Frecuencias de llegadas (a tiempo o con retraso) de vuelos de Marina y Julio.

Lo que ocurrió es que Marina voló, en sus **120** vuelos, una mayor cantidad de vuelos de día que de noche, y estas cantidades son **100** y **20**, respectivamente. Esto favorece el desempeño global de Marina, dado que es más factible llegar a tiempo de día que de noche. Lo contrario ocurrió con Julio, puesto que él voló la mayoría de sus vuelos de noche y no de día, siendo estas cantidades **100** y **20**, respectivamente.

Lo que vemos es que Marina está haciendo una afirmación equivocada al señalar que es mejor piloto que Julio, ya que no considera la relación entre el tiempo del aterrizaje y el momento en que este se realiza, día o noche. A este tipo de situación se le conoce como la *Paradoja de Simpson*.

3. Representaciones gráficas

Las representaciones gráficas constituyen una forma de organizar y presentar visualmente los datos. Existen diferentes tipos de gráficos que dependen del tipo de variable que representan. En el caso de las variables cualitativas, estudiaremos gráficos concretos, pictogramas, gráficos de barras y gráficos circulares. Estas representaciones pueden ser comprendidas por los alumnos desde los niveles más tempranos de enseñanza, aunque la complejidad, tanto en su construcción, donde vamos agregando elementos, como en la extracción de información que se puede hacer a partir de ellos, va aumentando en la medida que los alumnos adquieren mayores conocimientos y habilidades. En el caso de variables cuantitativas, estudiaremos diagramas de tallo y hojas, diagramas de puntos, histogramas, gráficos de línea o tendencia, y gráficos de dispersión. Algunas de estas representaciones pueden no estar en el currículo escolar de Educación Básica, sin embargo, son importantes para lograr un entendimiento global.

3.1. Gráficos concretos o reales

Los *gráficos concretos* o *reales* utilizan objetos para representar la frecuencia de cada categoría de una variable. Estos gráficos permiten a los alumnos manipular los objetos y visualizar la frecuencia de cada una de las categorías para, por ejemplo, verificar la diferencia entre la frecuencia de una categoría y otra.

Un gráfico concreto permite utilizar los objetos que los alumnos manipulan tanto en el aula como en sus entornos familiares. Un ejemplo de esto es el apilamiento de tarros de atún de diferentes marcas. En este caso, tarros de la misma marca se irán superponiendo, como se muestra en la **Figura III.3.** Cada marca de atún corresponde a una categoría y cada tarro apilado de una misma marca corresponde a una unidad observada de ella, de modo que el número de tarros representa una frecuencia absoluta.

Figura III.3: Gráfico concreto o real realizado a través del apilamiento de tarros de atún según su marca.

La **Figura III.4** muestra un gráfico concreto construido con fichas apiladas una sobre otra, agrupadas según su forma. La forma de las fichas (trapecio, triángulo, círculo y cuadrado) representa la categoría, y la cantidad de fichas apiladas, a la frecuencia absoluta. La **Figura III.5** muestra un gráfico real donde el color de los cubos indica la categoría y la cantidad de estos la frecuencia absoluta. En la figura se indica, además, el número que representa cada frecuencia de cada categoría, lo que permite relacionar la cantidad con el numeral.

Figura III.4: Gráfico concreto o real realizado a través del apilamiento fichas según su forma.

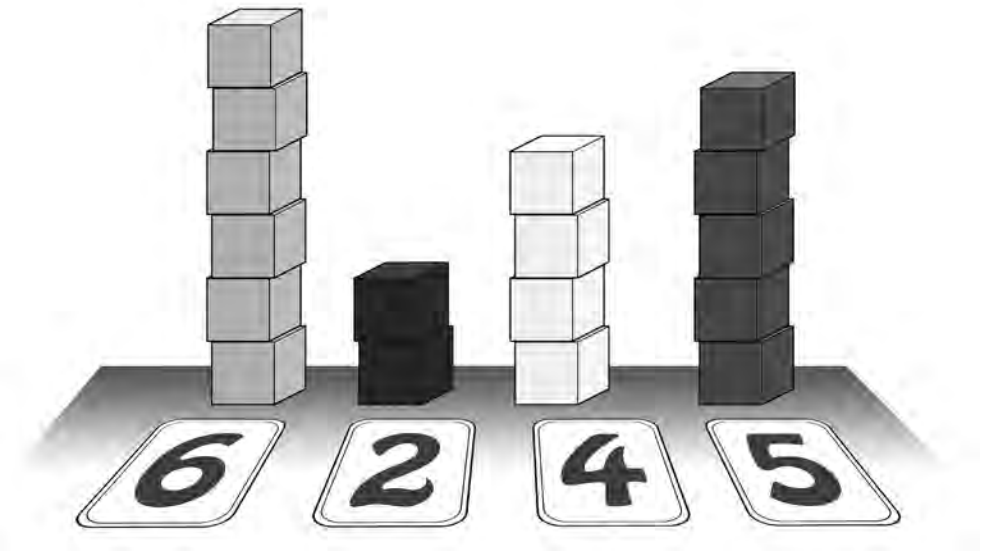


Figura III.5: Gráfico concreto o real que muestra, además, la frecuencia en cada una de las categorías.

Los gráficos reales corresponden a la primera aproximación que pueden hacer los niños en edad temprana al concepto de frecuencia. Por una parte, su construcción es relativamente simple y, por otra, entrega información visual sobre frecuencias, que también es relativamente sencilla de comprender. De este modo, la importancia de los gráficos concretos o reales radica en que corresponden al primer paso en la enseñanza de representaciones gráficas más abstractas.

En un gráfico real, es necesario identificar el atributo del objeto que determina la categoría de la variable a representar, la cual puede variar de acuerdo a la pregunta de interés y/o a la naturaleza del objeto. A modo de ejemplo, consideremos la **Figura III.4**. En ella, el atributo que determina las categorías corresponde a la forma de las fichas. Si queremos representar los deportes favoritos de los niños, cada forma, trapecio, triángulo, círculo y cuadrado, debe representar a uno de los deportes, fútbol, básquetbol, gimnasia rítmica y tenis. Cada una de las fichas apiladas representa una observación, en este caso, un niño.

El ámbito numérico en la construcción de gráficos concretos debe ir de **1** a **20**, dado que, más allá de estas cantidades, estos se vuelven ineficientes por el tiempo y la habilidad requerida para realizar los apilamientos. Cuando el número de observaciones es mayor, se requiere que el objeto que representa las observaciones reagrupe cantidades, como se observa en la **Figura III.6**. En la figura, cada una de las agrupaciones representa cinco unidades, requiriendo del conteo de **5** en **5** para determinar las frecuencias.

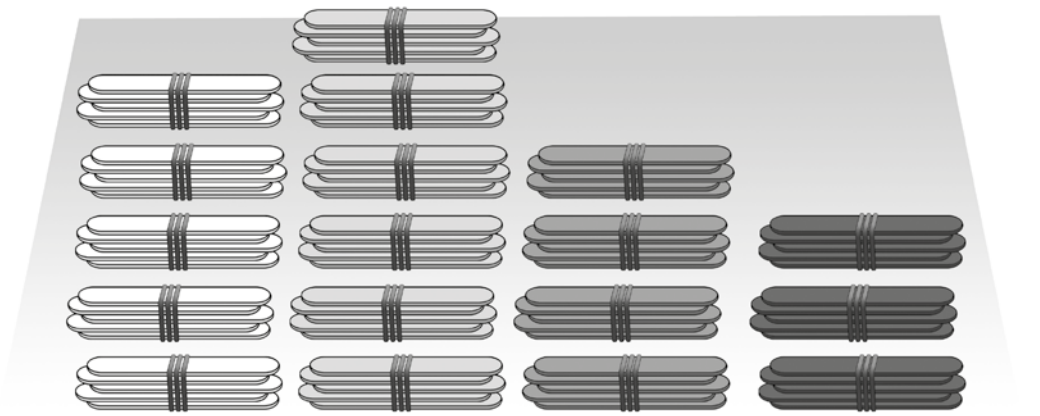


Figura III.6: Gráfico concreto donde objetos han sido agrupados para representar más de una unidad y facilitar el conteo.

En resumen, para construir un gráfico real se debe decidir qué objeto se utilizará para la representación de cada categoría, como, por ejemplo, cubos, e identificar el atributo del objeto que define las categorías, como, por ejemplo, el color del cubo. Luego, se deben representar las cantidades utilizando los objetos elegidos; apilando uno sobre otro, o bien yuxtaponiendo de forma ordenada, como en la **Figura III.6**, cada uno de ellos en la categoría correspondiente.

3.2. Pictogramas

3.2.1 Construcción e interpretación

Un *pictograma* corresponde a una representación gráfica más abstracta de un gráfico concreto o real, donde las frecuencias son representadas gráficamente a través de figuras. Un pictograma corresponde a un paso intermedio desde las representaciones concretas a representaciones abstractas, como los gráficos de barras que estudiaremos más adelante.

Un pictograma consiste en un gráfico en el cual se utilizan dibujos o símbolos para representar la frecuencia con la que se observa cada categoría, donde cada símbolo puede representar más de una unidad. Su principal utilidad es la comparación y el análisis de cantidades relacionadas a una misma variable, además de servir para resumir información, pudiendo incluso remplazar a una tabla de frecuencias.

El símbolo utilizado corresponde a una forma gráfica que indica un cierto número de observaciones, o frecuencia absoluta, y, por lo tanto, debe estar relacionado con la variable representada. La cantidad de observaciones en cada categoría se representa por la repetición del símbolo. Los símbolos utilizados pueden agregarse de manera horizontal, como en la **Figura III.7**, o de manera vertical.

La **Figura III.7** corresponde a un pictograma que muestra el número de árboles plantados en los años 2008, 2009 y 2010. Al pie del gráfico se observa que la figura utilizada, un árbol, representa **4.000** unidades plantadas. De este modo, podemos leer que, por ejemplo, en el año 2008 se plantaron **3 x 4.000 = 12.000** árboles. En la figura podemos distinguir un título y los nombres de las categorías de la variable representada. Estos elementos son necesarios en todo gráfico.

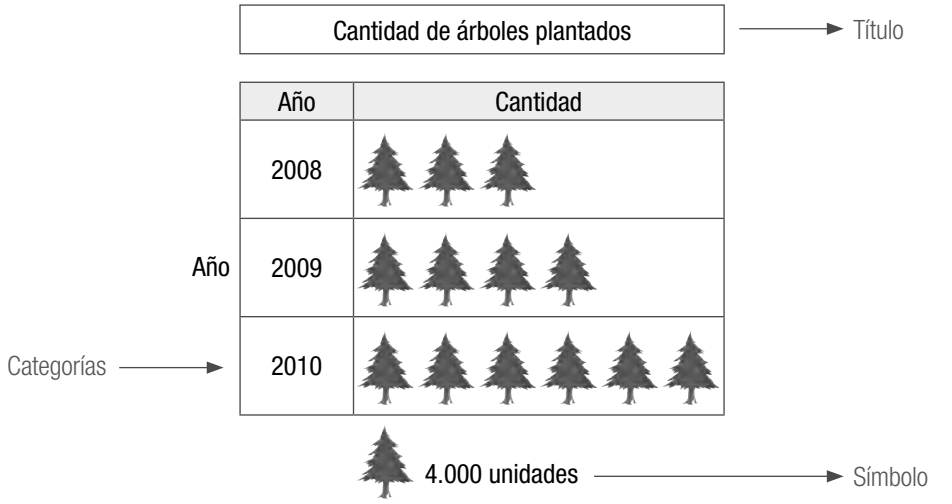


Figura III.7: Pictograma horizontal que representa la cantidad de árboles plantados en los años 2008, 2009 y 2010.

En el pictograma de la **Figura III.7**, la variable de interés es el año de plantación, que corresponde a una variable ordinal. Notamos entonces que en la figura las categorías, 2008, 2009 y 2010, deben ser ordenadas.

En la enseñanza de pictogramas, es recomendable que inicialmente el símbolo utilizado represente solo una unidad. En niveles más avanzados es posible utilizar símbolos que representen un mayor número de observaciones, como en el pictograma en la **Figura III.7**, en cuyo caso se debe tener en consideración el ámbito numérico adecuado al nivel de conocimiento de los niños.

También es importante considerar que existen símbolos que pueden ser divisibles para representar una porción de lo que representa el símbolo completo, como, por ejemplo, barras de chocolate. Sin embargo, en este contexto, para un niño no es razonable que se dividan símbolos que en la realidad no son divisibles, como, por ejemplo, las personas. Si bien se entiende el concepto que hay detrás de graficar "media persona", para un niño pequeño esta representación es incomprensible.

Es bastante frecuente encontrar, tanto en medios de comunicación como en otras fuentes, pictogramas donde las figuras no representan el número de observaciones por repetición de un símbolo, sino por el tamaño de este, como en el pictograma en la **Figura III.8**.

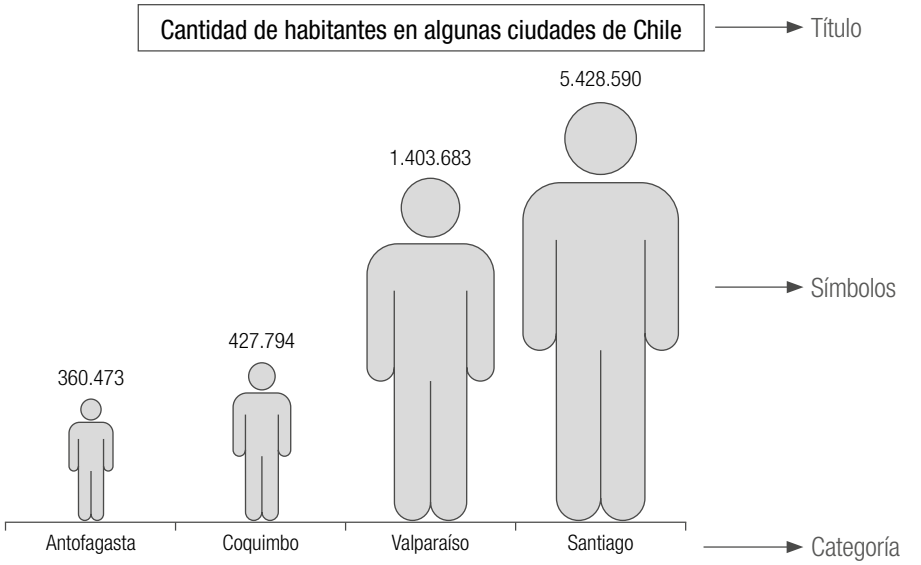


Figura III.8: Pictograma vertical que representa el número de habitantes en algunas ciudades de Chile. Las frecuencias absolutas se representan gráficamente a través del tamaño de la figura. No es recomendable utilizar esta opción.

En el pictograma de la **Figura III.8**, símbolos de mayor tamaño han sido utilizados para mostrar un mayor número de habitantes. Sin embargo, ¿a qué nos referimos con "tamaño"? Tamaño puede ser representado tanto por la altura de la figura, como por su área, entre otras características. En cualquier caso, puede ocurrir que el criterio utilizado distorsione visualmente la información y comunique mensajes erróneos. En efecto, es posible que, aunque la figura indique el número exacto de observaciones en cada categoría, como en la **Figura III.8**, la impresión visual nos dé una noción equivocada de la realidad.

Para construir un pictograma se propone, primero, elegir el símbolo que se utilizará para la representación y el valor que tendrá. Luego, indicar en uno de los ejes, vertical u horizontal, las categorías que serán consideradas, y representar las cantidades utilizando el símbolo elegido a través del número de veces que se repite. Se debe también indicar el número de observaciones que representa cada símbolo, e incluir un título sucinto, pero informativo.

3.2.2 Errores en la construcción de pictogramas

Uno de los errores en la construcción de pictogramas corresponde a la utilización de símbolos de diferentes tamaños asociados a diferentes categorías. A modo de ejemplo, consideremos el pictograma en la **Figura III.9**, que muestra las mascotas de un grupo de niños. El pictograma utiliza símbolos diferentes para cada categoría, los que, al ser de diferentes tamaños, distorsionan la figura. En efecto, a modo de ejemplo, hay solo 3 niños cuya mascota es un conejo, sin embargo, por ser esta figura de mayor tamaño, la fila correspondiente a esta categoría es de mayor longitud, lo que distorsiona la información que se puede extraer del pictograma.

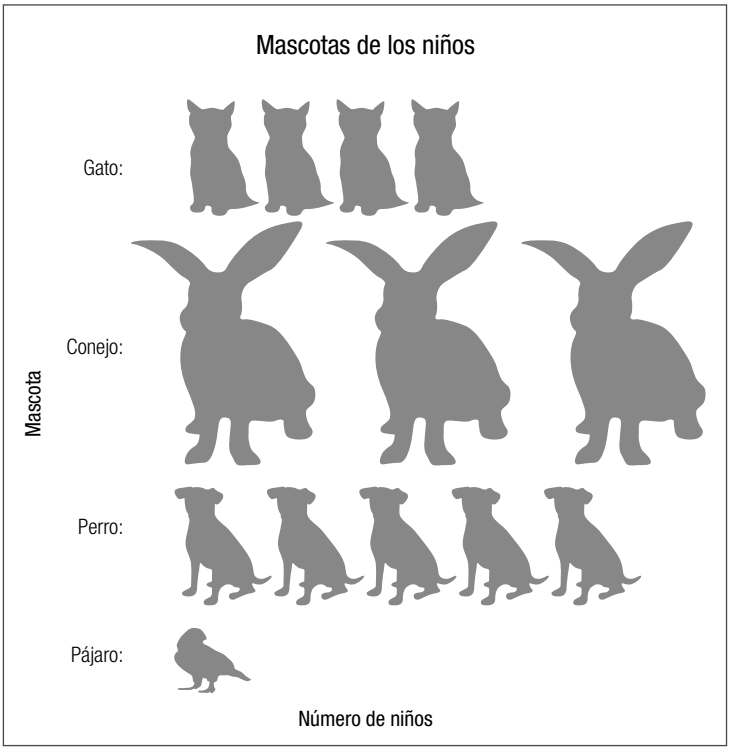
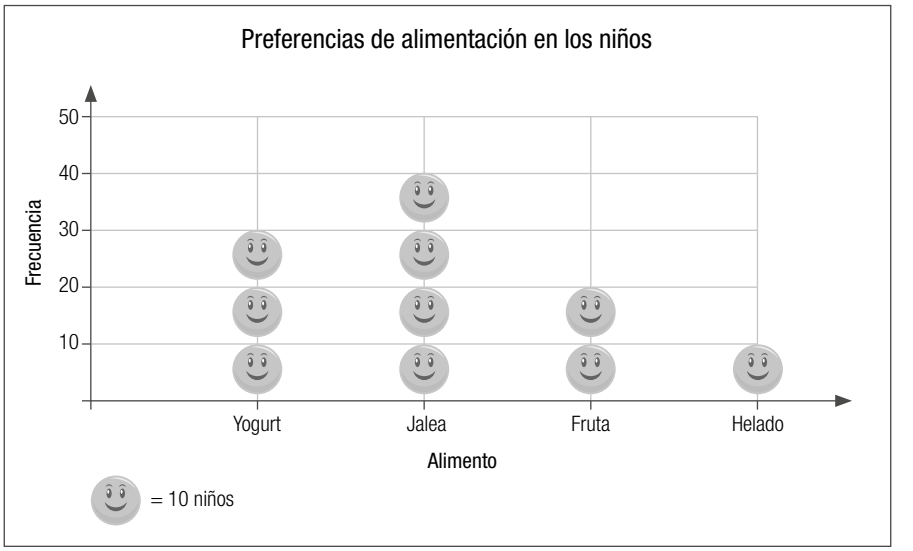


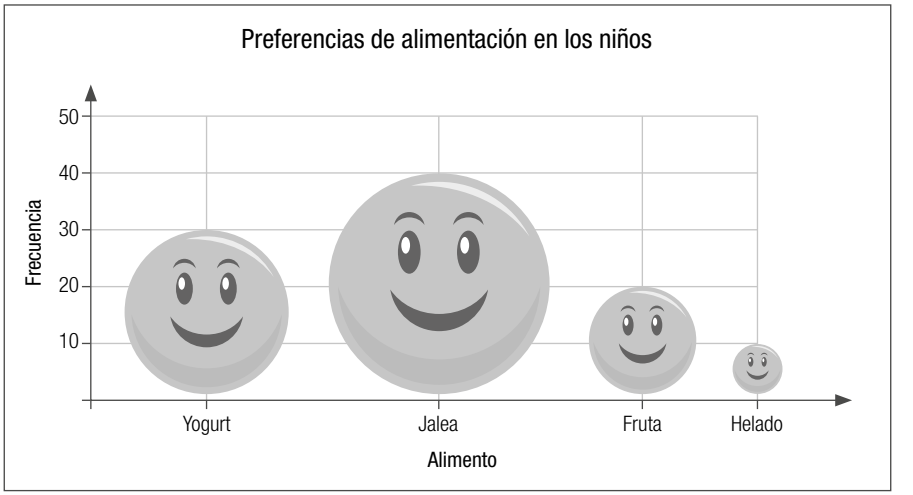
Figura III.9: Pictograma que utiliza símbolos de diferentes tamaños, lo que distorsiona la información entregada por las observaciones.

Este error corresponde a un error de *origen comunicativo*, dado que la representación visual no muestra, o distorsiona, las obligaciones.

Ejercicio

Considere los dos pictogramas en las figuras siguientes, referentes a la preferencia de alimentos de **100** niños entrevistados.





a. ¿Entregan ambos pictogramas la misma información objetiva sobre el número de niños que prefieren cada una de las alternativas?

b. ¿Entregan ambos pictogramas la misma información visual sobre el número de niños que prefieren cada una de las alternativas?

c. ¿Qué pictograma preferiría? ¿por qué?

d. ¿Cuál es la tabla de frecuencias correspondiente?

3.3. Gráficos de barras

3.3.1 Gráficos de barras simples

Un *gráfico de barras* corresponde a una representación más abstracta de un pictograma. En él, los símbolos en una misma categoría han sido integrados en una barra o rectángulo, obteniéndose tantas barras paralelas como número de categorías de la variable representada. La **Figura III.10** muestra un gráfico de barras construido en base a la información entregada en la "5ta encuesta nacional de la juventud", sobre las regiones a las que pertenece la población joven. Podemos leer, por ejemplo, que aproximadamente **1.600.000** jóvenes pertenecen a la Región Metropolitana.

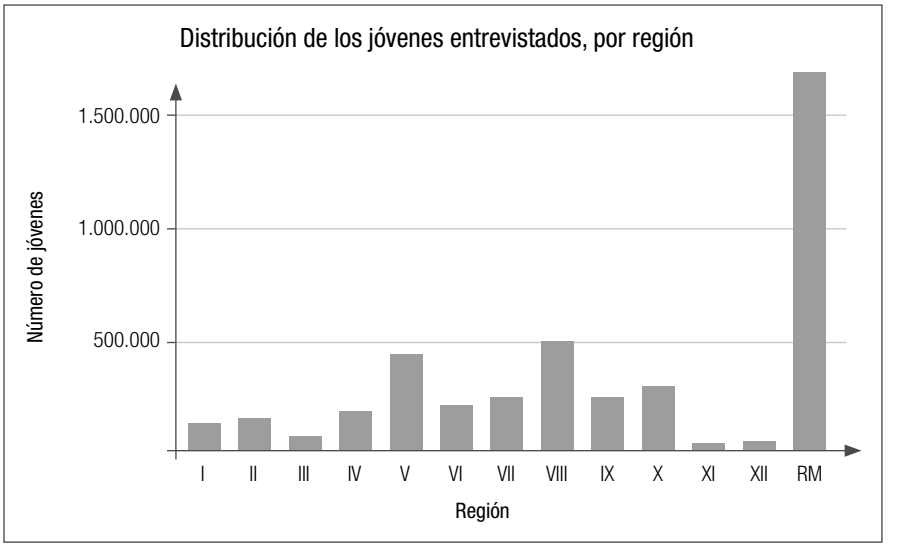


Figura III.10: Gráfico de barras de la distribución de las regiones a las que pertenece la población joven. Se muestran las frecuencias absolutas.

Con este tipo de gráficos, se deben utilizar barras con la misma base, como en la **Figura III.10**. El alto de las barras corresponde a la frecuencia absoluta de la categoría representada. Sin embargo, también es posible representar la frecuencia relativa porcentual como en la **Figura III.11**. En los dos casos, la forma de la figura permanece intacta, por lo que nos entrega la misma información cualitativa. Las barras deben ser presentadas separadas y equidistantes unas de otras.

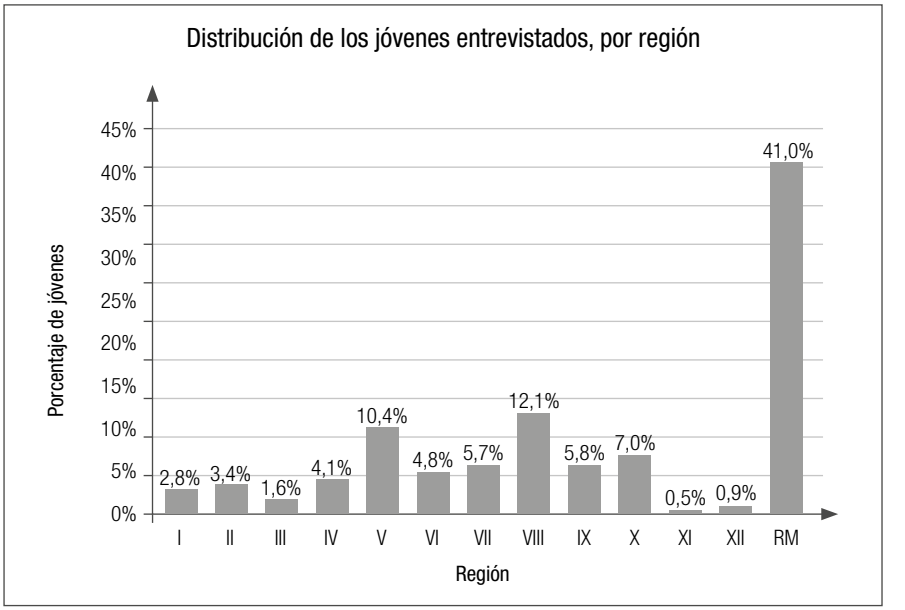


Figura III.11: Gráfico de barras de la distribución de las regiones a las que pertenece la población joven. Se muestran las frecuencias relativas porcentuales.

Tanto en la **Figura III.10** como en la **III.11**, notamos que los jóvenes provienen principalmente de la Región Metropolitana, ya que su porcentaje es más de tres veces mayor que los de las regiones que le siguen, V y VIII. También observamos que los porcentajes más bajos de jóvenes provienen de las regiones XI y XII.

Si bien el eje vertical de un gráfico de barras nos permite leer aproximadamente el alto de las barras, puede resultar útil entregar estas cantidades sobre las barras, ya sea como frecuencias o como porcentajes, según las unidades indicadas en el eje cuantitativo, como lo hace la **Figura III.11**.

Al igual que todo gráfico, el gráfico de barras debe llevar un título sucinto, pero informativo. Se distingue, además, un eje cualitativo en el que se disponen las categorías de la variable que se quiere representar. Usualmente, este eje corresponde al eje horizontal o de las abscisas, como en las **Figuras III.10** y **III.11**. Sin embargo, al igual que en el caso de los pictogramas, es posible utilizar el eje vertical o de las ordenadas como eje de las categorías, obteniendo barras horizontales, como en el gráfico que se muestra en la **Figura III.12**.

Los gráficos de barras son útiles para todo tipo de variables cualitativas, ya sean nominales u ordinales. En el caso de variables ordinales, la figura debe respetar el orden de las categorías en el eje cualitativo. De este modo, será posible analizar la evolución de las frecuencias de las observaciones al movernos a través de las categorías.

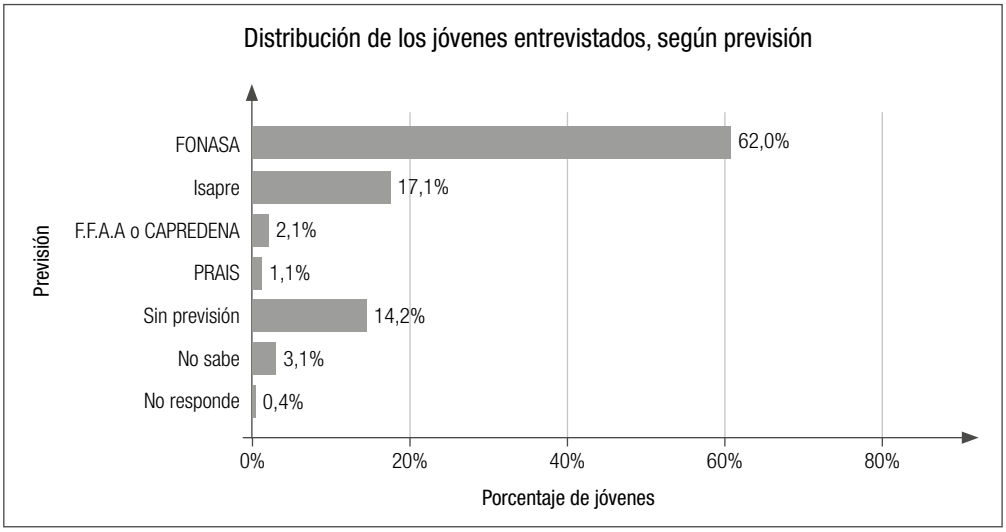


Figura III.12: Gráfico de barras horizontal de la distribución de la previsión social a la que pertenece la población joven.

Para construir un gráfico de barras, se deben obtener previamente las frecuencias absolutas o relativas porcentuales de las categorías a graficar. Esta información puede ser extraída a partir de una tabla de frecuencias. A modo de ejemplo, la **Figura III.12** puede ser construida a partir de la tabla de frecuencias en la **Tabla III.35**[15].

Previsión	Porcentaje de jóvenes
FONASA	62,0%
Isapre	17,1%
F.F.A.A. o CAPREDENA	2,1%
PRAIS	1,1%
Sin previsión	14,2%
No sabe	3,1%
No responde	0,4%

Tabla III.35: Previsión social a la que pertenece la población joven.

Luego, se deben ubicar, separadas y equidistantes, las categorías a graficar en el eje cualitativo, construir el eje cuantitativo con las unidades de las frecuencias absolutas o relativas porcentuales, según lo que se haya decidido graficar, construir un rectángulo de igual base para cada categoría hasta la altura, en el caso de un gráfico vertical, o ancho, en un gráfico horizontal, de su frecuencia en el eje cuantitativo. Finalmente, se debe incluir un título.

Los gráficos de barras mostrados en las **Figura III.10**, **III.11** y **III.12** se denominan *gráficos de barras simples*, puesto que representan una sola variable, en ese caso, la región de pertenencia o la previsión social.

[15] Los porcentajes no suman 100% debido al redondeo.

3.3.2 Gráficos de barras agrupadas

Al igual que con tablas de frecuencias de dos entradas, los gráficos de barras pueden ser utilizados para mostrar la relación entre dos variables categóricas. En estos casos, se utilizan *gráficos de barras agrupadas*.

Trabajaremos en base a un ejemplo, retomando los resultados de la "Encuesta nacional de hábitos de actividad física y deportes en la población chilena de 18 años y más"[16] del año 2012. Como vimos en la **Sección 2** sobre tablas de frecuencias, en esta encuesta se preguntó a los entrevistados por su interés en el deporte y la actividad física, y su práctica. Esta información fue desglosada según sexo. El gráfico en la **Figura III.13** muestra los porcentajes de entrevistados en cada situación de interés o desinterés por el deporte y la actividad física, y práctica, separados por sexo. Este tipo de gráfico corresponde a un gráfico de barras agrupadas.

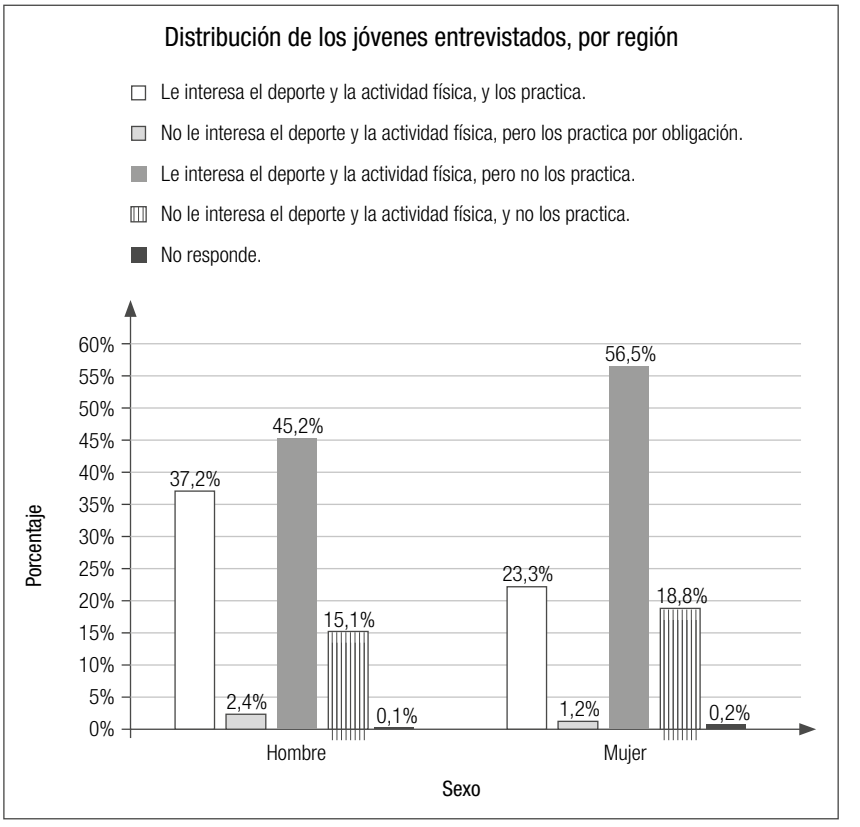


Figura III.13: Gráfico de barras agrupadas que muestra la distribución del interés en el deporte y la actividad física y su práctica, para cada sexo.

En el gráfico leemos, por ejemplo, que a un **37,2%** de los hombres entrevistados les interesa el deporte y la actividad física, y los practican. También leemos que este porcentaje baja a **23,3%** en las mujeres. Sin embargo, a un mayor porcentaje de mujeres que hombres (**56,5%** versus **45,2%**) les interesa el deporte y la actividad física, pero no los practican. ¿Puede aventurar explicaciones?

[16] Información tomada de http://www.ind.cl/estudios-e-investigacion/investigaciones/Documents/2012/Encuesta%20Act%20Fisica/encuesta-act-fisica-2012.pdf

Para construir un gráfico como el de la **Figura III.13**, resulta útil obtener previamente una tabla de frecuencias de dos entradas condicional. En el ejemplo, requerimos de una tabla de frecuencias relativas porcentuales que muestre el porcentaje de entrevistados en cada situación de interés o desinterés por el deporte y la actividad física, y su práctica, para cada sexo por separado. Esta tabla fue construida en la **Sección 2**, y la replicamos aquí para referencia, como **Tabla III.36**.

Interés y práctica	Sexo	
	Masculino	Femenino
Le interesa el deporte y la actividad física, y los practica.	37,2%	23,3%
No le interesa el deporte y la actividad física, pero los practica por obligación.	2,4%	1,2%
Le interesa el deporte y la actividad física, pero no los practica.	45,2%	56,5%
No le interesa el deporte y la actividad física, y no los practica.	15,1%	18,8%
No sabe o no responde.	0,1%	0,2%
Total	**100%**	**100%**

Tabla III.36: Tabla de frecuencias relativas porcentuales condicional del interés en la actividad física y el deporte y su práctica, para cada sexo de los entrevistados.

Notamos que los porcentajes en la primera columna de la **Tabla III.36**, asociados a los hombres, se muestran en el grupo de barras en la izquierda del gráfico de barras agrupadas en la **Figura III.13**. Del mismo modo, los porcentajes en la segunda columna de la misma tabla, asociados a las mujeres, se muestran en el grupo de barras en la derecha del gráfico de barras agrupadas en la **Figura III.13**.

Notamos que la suma de los porcentajes en las barras dentro de cada sexo, o en los dos grupos de barras, corresponde a **100%**.

La mayor ventaja de un gráfico de barras agrupado radica en la facilidad de cruzar la información entregada por dos variables de interés. Como hemos visto en este ejemplo, podemos visualizar la relación entre el interés en el deporte y la actividad física y su práctica, y el sexo de los entrevistados, lo que no podríamos hacer a partir de los gráficos de barras simples de cada variable por separado.

3.3.3 Errores en la construcción de gráficos de barras

Tal como vimos en la construcción de pictogramas, también pueden ocurrir errores de *origen comunicativo* en la construcción de gráficos de barras. A modo de ejemplo, consideremos el gráfico en la **Figura III.14**, que muestra el número de discos vendidos en una disquería entre los años 2006 y 2009. La figura presenta varias dificultades de lectura. La primera, producida por la utilización de una perspectiva inapropiada. En efecto, no utilizar una vista frontal dificulta la comparación de frecuencias. Por otra parte, la figura utiliza una representación en tres dimensiones. El ancho de las barras utilizadas presenta otra dificultad en la comparación de frecuencias a través de las barras. Por último, la figura ordena las categorías de forma inversa al orden natural, al localizarlas en orden decreciente, desde 2009 a 2006.

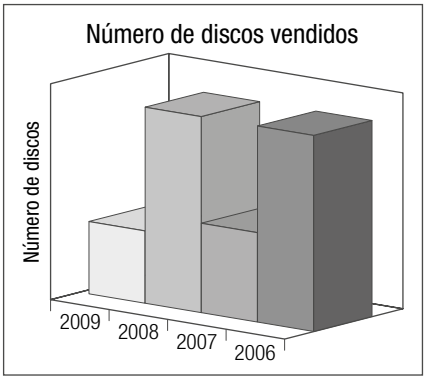


Figura III.14: Gráfico de barras en tres dimensiones. El gráfico distorsiona la información.

En otro ejemplo de un error de origen comunicativo, la **Figura III.15** muestra dos gráficos de barras diferentes, ambos construidos utilizando el mismo conjunto de observaciones sobre el número de alumnos matriculados en una escuela. En el gráfico de la izquierda, el eje vertical, o de las ordenadas, no comienza en el origen y entrega visualmente la impresión de una gran diferencia entre los números de alumnos matriculados en los años 2010 y 2011. Sin embargo, la figura de la derecha entrega la información adecuada, al comenzar el eje vertical en el origen, mostrando que la diferencia entre los números de alumnos matriculados en los dos años es más pequeña, en comparación con la magnitud de las cantidades consideradas.

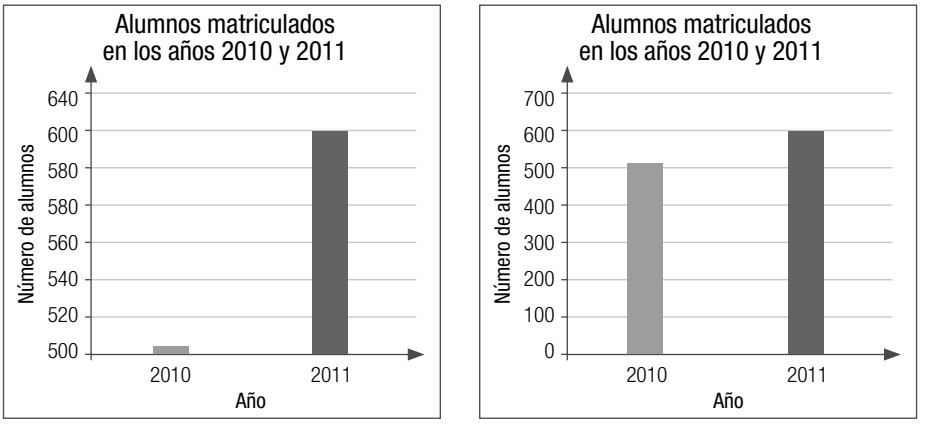


Figura III.15: Gráficos de barras del número de alumnos matriculados en una escuela. El gráfico de la izquierda utiliza una escala inapropiada, lo que distorsiona la información.

Ejercicio

En un curso de **39** alumnos se recolecta información sobre la comuna en que viven los niños y su sexo, con el objetivo de conocer la posible asociación entre sexo y comuna. Considere los gráficos de barras agrupadas que se muestran en las **Figuras a.** y **b.**

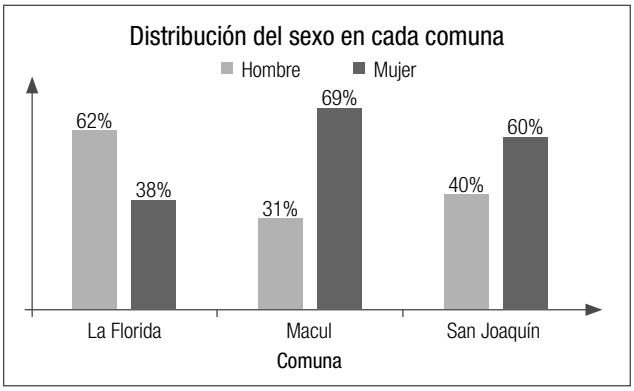


Figura a: Distribución del sexo de los niños, en cada comuna.

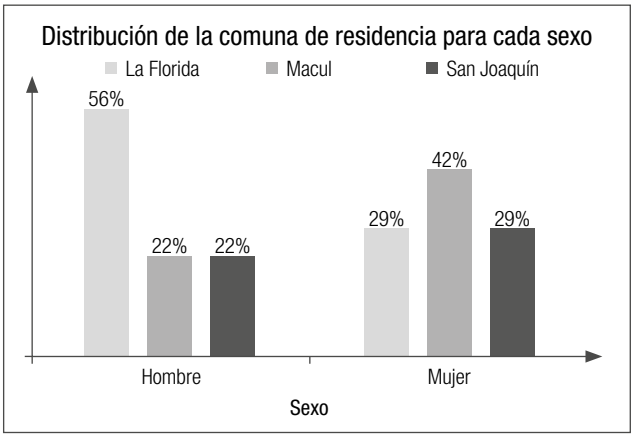


Figura b: Distribución de la comuna en que viven los niños, según sexo.

Si bien ambos gráficos de barras son similares, ellos comunican diferente información, la que puede identificarse a través del título de cada uno de ellos.

a. ¿A partir de qué gráfico de barras puede usted determinar el porcentaje de mujeres en la comuna de Macul? ¿a qué porcentaje corresponde?
b. Si consideramos solo las mujeres, ¿a partir de qué gráfico puede usted determinar el porcentaje de las que viven en la comuna de San Joaquín? ¿a qué porcentaje corresponde?
c. ¿Cuál es el porcentaje de mujeres en la comuna de La Florida?
d. ¿En qué comuna vive la mayor cantidad de hombres?
e. ¿Aproximadamente, qué porcentaje de hombres hay en esa comuna?

Es frecuente encontrar gráficos o figuras que representan información estadística utilizando barras y que, sin embargo, no corresponden al tipo de gráficos de barras que aquí hemos discutido. Consideremos, a modo de ejemplo, el gráfico en la **Figura III.16**.

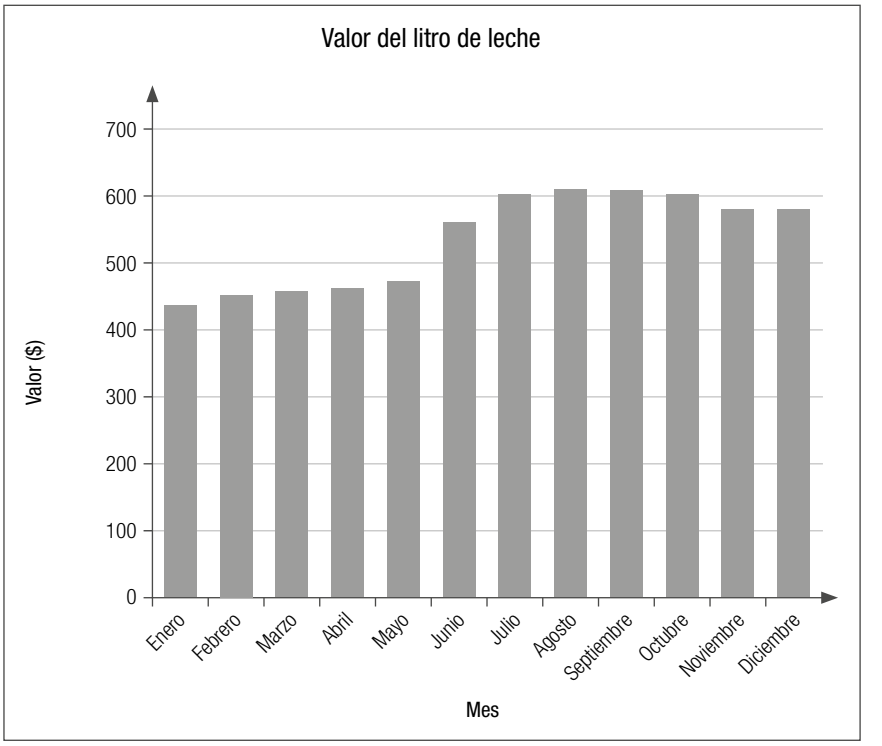


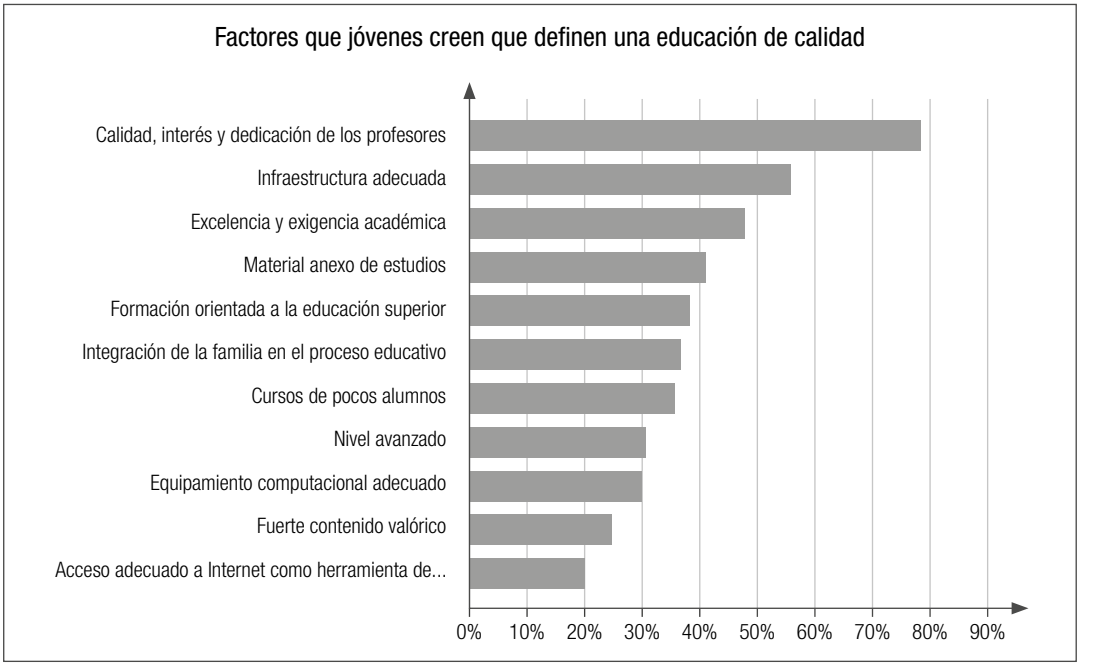
Figura III.16: Valor del litro de leche, en pesos, desde enero a diciembre de un mismo año.

La **Figura III.16** muestra el precio del litro de leche, expresado en pesos, desde enero a diciembre de un mismo año. La altura de la barra indica el valor promedio del litro de leche durante dicho mes. A modo de ejemplo, la figura indica que en marzo de dicho año, el valor del litro de leche fue de, aproximadamente, **$450**. Notamos que el eje vertical, o de las ordenadas, no representa una frecuencia (absoluta o relativa), sino un valor monetario. Luego, si bien el gráfico utiliza barras para entregar la información deseada, no corresponde a un gráfico de barras en el contexto que aquí hemos utilizado.

Aunque la **Figura III.16** no puede considerarse incorrecta, existen otros tipos de representaciones gráficas más adecuados para situaciones como esta, por ejemplo, los gráficos de línea o tendencia, que estudiaremos más adelante.

Ejercicio

Considere el siguiente gráfico que muestra los factores que los jóvenes creen que definen una Educación Básica y Media de calidad.



Si bien ambos gráficos de barras son similares, ellos comunican diferente información, la que puede identificarse a través del título en cada uno de ellos.

a. ¿Corresponde la figura a un gráfico de barras en el sentido en que lo hemos aprendido? ¿por qué?
b. ¿Cómo cree usted que fue redactada la pregunta?
c. ¿Cuáles son los principales factores percibidos como definitorios de una educación de calidad?

3.4. Gráficos circulares o de torta

3.4.1 Gráficos de barras simples

Un *gráfico circular* o *de torta* se utiliza para representar las frecuencias relativas porcentuales de cada categoría de la variable de interés. Los datos se representan mediante sectores de un círculo, cuyos ángulos o áreas son proporcionales a las frecuencias de las categorías que representan. De este modo, un gráfico circular permite comparar fácilmente los porcentajes de cada una de las categorías de interés, a través de la comparación visual de las áreas de las secciones.

La **Figura III.17** corresponde a un gráfico circular que muestra la distribución de la población joven según nivel socioeconómico entregada por la "5ta encuesta nacional de la juventud". En él podemos ver que el mayor porcentaje de la población joven pertenece al nivel socioeconómico C3, correspondiente a un **33,4%**, mientras que el nivel E es el que tiene menor representación, con un **8,1%**. Notamos que, a mayor porcentaje, se ha asociado una porción de mayor área o, equivalentemente, de mayor ángulo que la contiene.

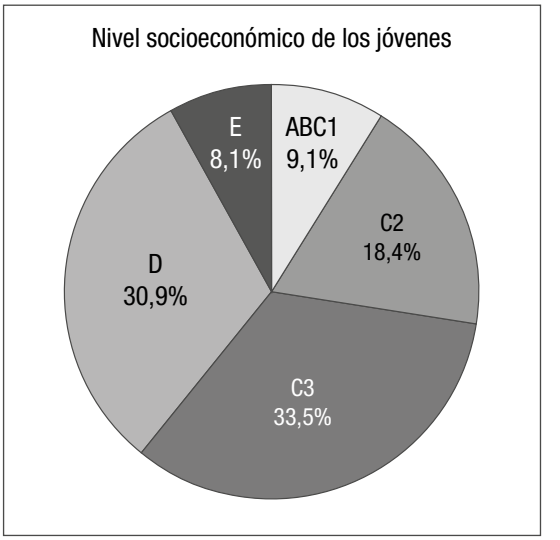


Figura III.17: Gráfico circular o de torta de la distribución del nivel socioeconómico en la población joven.

En la **Figura III.17**, la leyenda al interior de cada sector circular indica la categoría que representa y el porcentaje de dicha categoría dentro del conjunto de datos. Una manera alternativa de entregar esta información es presentar una leyenda fuera del círculo, que identifique la categoría con el sector que la representa, a través de color, achurado u otra característica visual, como en la **Figura III.18**.

En cada tipo de representación gráfica que hemos estudiado, nos hemos detenido en describir su construcción de manera detallada; sin embargo, hemos hecho hincapié en la importancia de la lectura e interpretación del gráfico. Esto es especialmente cierto en el caso de los gráficos circulares. En efecto, los alumnos son capaces de leer e interpretar gráficos circulares mucho antes de entender proporciones o ángulos para poder construirlos. Es por esto que resulta posible posponer la enseñanza de la construcción a niveles más avanzados, presentando la figura y su interpretación en niveles anteriores.

Para construir un gráfico circular, se deben obtener las frecuencias relativas porcentuales de las categorías a graficar, y utilizarlas para obtener la medida de los ángulos de los sectores circulares, estableciendo una relación entre las primeras y los 360° del círculo completo.

A modo de ejemplo, retomemos el problema sobre la preferencia de ciertos ingredientes de pizza, cuyas frecuencias se muestran nuevamente en la **Tabla III.37**.

Ingrediente	Número de entrevistados
Anchoas	8
Queso	27
Peperoni	16
Vienesas	36
Vegetales	23
Total	**110**

Tabla III.37: Ingredientes de pizza preferidos por los entrevistados.

Agregaremos, en la **Tabla III.38**, las columnas de frecuencias relativas y ángulos de los sectores que representarán a cada categoría. A modo de ejemplo, la frecuencia relativa porcentual de la categoría "Anchoas" es 7%, por lo que el ángulo del sector que representará a esta categoría corresponde a $\frac{7}{100}$ x 360º, o aproximadamente 25º. La frecuencia relativa porcentual de la categoría "Queso" es 25%, por lo que el ángulo del sector que representará a esta categoría corresponde a $\frac{25}{100}$ x 360º = 90º. De la misma forma, se completan los valores restantes de la tabla. Notamos que la suma total de los ángulos de los sectores debe ser igual a 360º.

Ingrediente	Número de entrevistados	Frecuencia relativa porcentual	Ángulo (grados)
Anchoas	8	7%	25°
Queso	27	25%	90°
Peperoni	16	15%	54°
Vienesas	36	32%	115°
Vegetales	23	21%	76°
Total	**110**	**100%**	**360°**

Tabla III.38: Ingredientes de pizza preferidos por los entrevistados. Se indican los ángulos de los sectores que representarán a las categorías en un gráfico de torta o circular.

Para representar la información entregada en la **Tabla III.38** en un gráfico circular o de torta, dibujamos un círculo y construimos las secciones según los ángulos en la última columna de la **Tabla III.38**, utilizando un transportador. Mostramos las frecuencias relativas porcentuales al interior de cada sector y construimos una leyenda para identificar qué categoría representa cada parte del círculo. Finalmente agregamos un título adecuado. El gráfico circular obtenido en este caso se muestra en la **Figura III.18**. Enfatizamos que la figura no debe indicar el tamaño del ángulo del sector, sino que la frecuencia relativa porcentual que este representa.

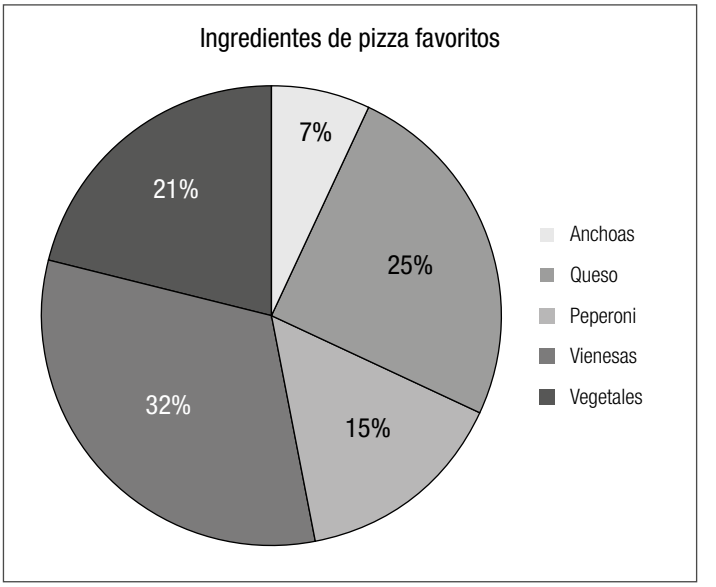


Figura III.18: Gráfico circular o de torta que representa las frecuencias relativas porcentuales de los ingredientes de pizza preferidos de los entrevistados.

Recalcamos que la construcción que hemos seguido tiene fines principalmente ilustrativos, dado que en la práctica se cuenta con herramientas computacionales de fácil acceso que entregan este tipo de gráficos. Esto es lo que se ha utilizado para la **Figura III.18**.

3.4.2 Errores en el uso de gráficos circulares

Una de las dificultades de *origen matemático* es la falta de atención a las partes y el todo cuando se decide utilizar el gráfico circular. El gráfico circular solo es apropiado cuando la frecuencia o el porcentaje de las observaciones dentro de cada categoría forman las partes de un todo.

La **Tabla III.39** muestra los porcentajes de personas en un grupo que compraron música a través de un sitio de Internet, durante el mes pasado, por cada grupo de edad.

Edad por grupo	Porcentaje que compró música en línea
De 12 a 17 años	24%
De 18 a 24 años	21%
De 25 a 34 años	20%
De 35 a 44 años	16%
De 45 a 54 años	10%
De 55 a 64 años	3%
De 65 años o mas	1%

Tabla III.39: Porcentajes de personas que compraron música a través de un sitio de Internet durante el mes pasado, dentro de cada grupo de edad.

A modo de ejemplo, la **Tabla III.39** indica que, dentro del grupo de personas de 12 a 17 años, el 24% compró música en línea durante el mes anterior.

Un error corresponde a utilizar un gráfico circular en una situación como esta, únicamente guiado por el hecho de que la información es entregada en forma de porcentajes asociados a categorías. Sin embargo, notemos que cada porcentaje que se muestra fue obtenido sobre un grupo de personas diferente. A modo de ejemplo, en las penúltima y última fila leemos los porcentajes 3% y 1%. Cada uno de estos porcentajes fue obtenido sobre un total de observaciones o individuos diferentes. En el primero caso, el valor 3% fue obtenido sobre el total de personas de 55 a 64 años, mientras que, en el segundo caso, el valor 1% fue obtenido sobre el total de personas de 65 años o más. En estas situaciones, no es adecuado construir un gráfico circular, dado que el círculo utilizado en un gráfico circular representa el todo, o grupo único, con respecto al cual fueron obtenidos los porcentajes de las categorías. En el ejemplo, los porcentajes no fueron obtenidos con respecto a un único grupo, sino a cada grupo de edad, por separado.

En otro ejemplo, consideremos la **Tabla III.40**, que muestra los porcentajes de niños y niñas en tres cursos paralelos de una escuela.

Curso	Porcentaje de niños	Porcentaje de niñas
Cuarto A	51,1%	48,9%
Cuarto B	51,4%	48,6%
Cuarto C	50,7%	49,3%

Tabla III.40: Porcentajes de niños y niñas dentro de cada curso.

En este caso, el mismo error que se describe en el caso anterior corresponde a la construcción de un gráfico circular para niños y para niñas, por separado, como los que se muestran en la **Figura III.19.** A modo de ejemplo, los tres porcentajes asociados a los niños, 51,1%, 51,4% y 50,7%, fueron obtenidos en base a grupos de niños diferentes, cuarto A, cuarto B y cuarto C, respectivamente; luego, no deben ser representados en un mismo círculo. Lo mismo es válido para las niñas.

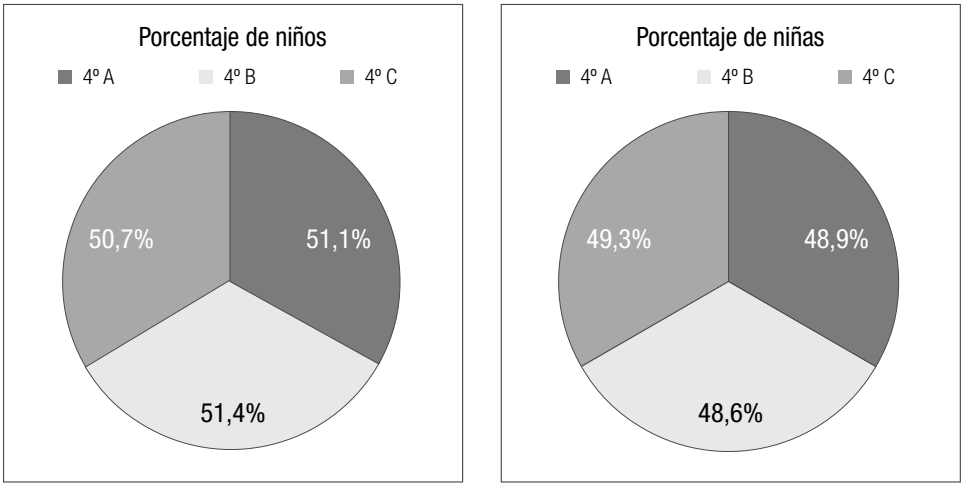


Figura III.19: Gráficos circulares erróneos. Dentro de cada círculo, los porcentajes fueron obtenidos sobre grupos de niños diferentes (cuarto A, cuarto B y cuarto C).

Por otra parte, notemos, a modo de ejemplo, que los dos porcentajes en la primera fila, **51,1%** y **48,9%**, fueron obtenidos sobre el mismo grupo de niños, el cuarto A. Luego, podemos representar dichos porcentajes en un gráfico circular, repitiendo luego para los cuartos B y C. En este caso, se obtiene la **Figura III.20**.

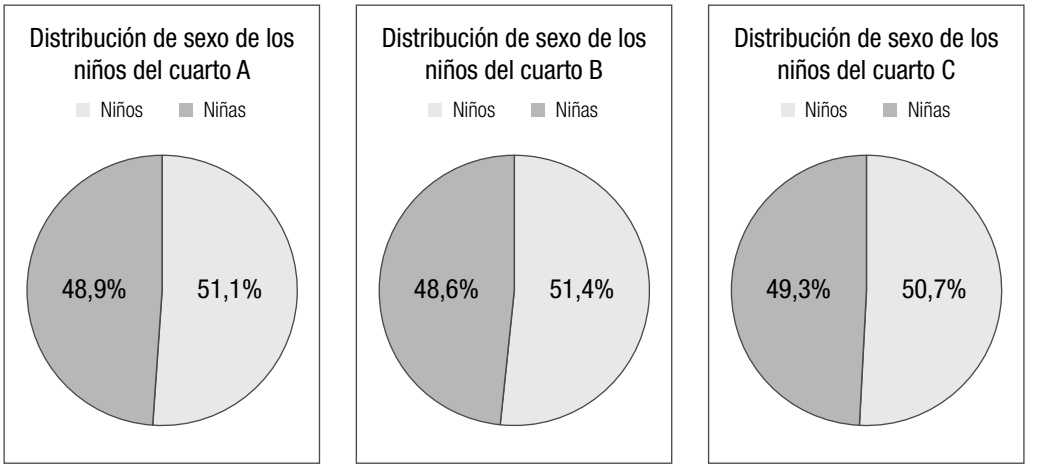


Figura III.20: Gráficos circulares correctos. Dentro de cada círculo, los porcentajes fueron obtenidos en el mismo grupo de niños (izquierda: cuarto A, centro: cuarto B, derecha: cuarto C).

Está claro que cuando los porcentajes considerados no suman **100%**, no fueron obtenidos sobre un mismo grupo, por lo que no deben ser representados en un mismo gráfico circular. Sin embargo, el hecho que un grupo de porcentajes sume **100%** no es motivo suficiente para hacerlo: se debe verificar que, efectivamente, hayan sido obtenidos sobre un mismo grupo de observaciones.

En resumen

- Un *gráfico concreto* o *real* es una representación concreta de las frecuencias de las observaciones en cada categoría. Cada unidad apilada o yuxtapuesta representa una observación en la categoría correspondiente.
- Un *pictograma* es un gráfico que utiliza dibujos o símbolos para representar la frecuencia de las categorías de una variable. Cada símbolo puede representar más de una observación.
- Un *gráfico de barras* es una representación abstracta de las frecuencias de las categorías de una variable. En él, los símbolos son integrados en una barra o rectángulo.
- Un *gráfico de barras agrupadas* es un gráfico de barras para dos variables, que muestra la asociación que puede existir entre estas.
- Un *gráfico circular* o *de torta* representa la frecuencia relativa porcentual de las categorías de una variable en sectores circulares.

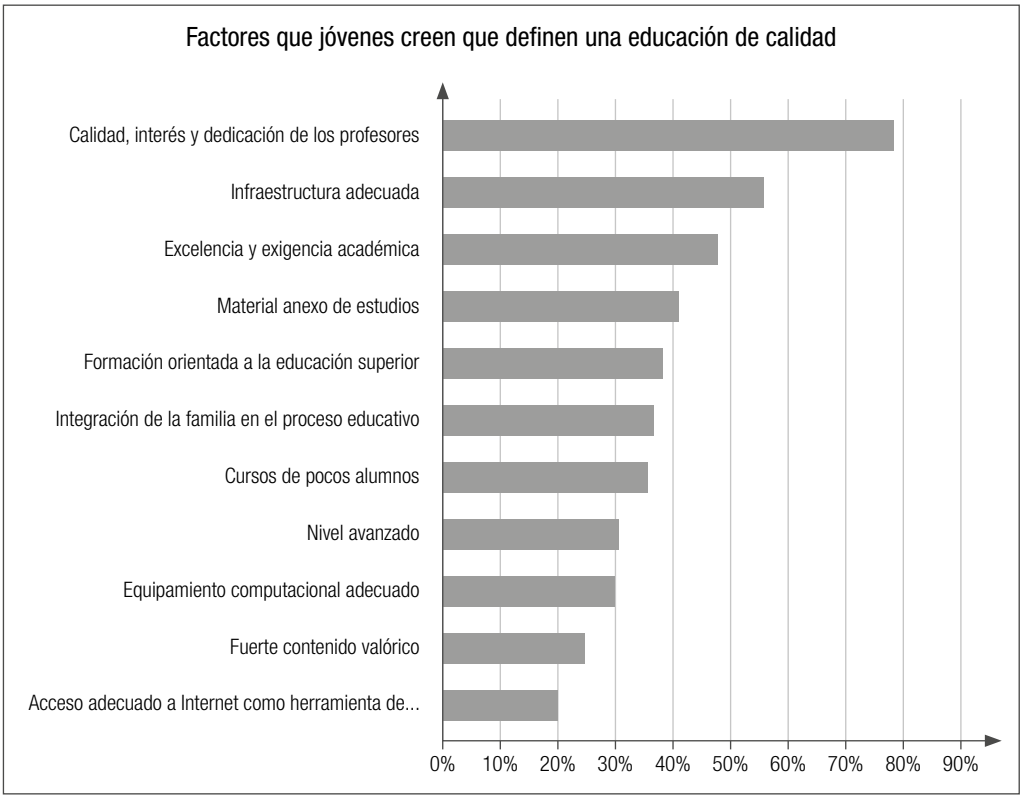
Ejercicios

1. Suponga que se pide a los niños del curso que expresen su opinión sobre la afirmación: "Me gusta jugar con mis amigos/amigas durante los recreos", y les dan las alternativas: En desacuerdo - Ni de acuerdo ni en desacuerdo - De acuerdo. Se obtienen los siguientes resultados:

Grado de acuerdo	Porcentaje
En desacuerdo	5%
Ni de acuerdo ni en desacuerdo	40%
De acuerdo	55%

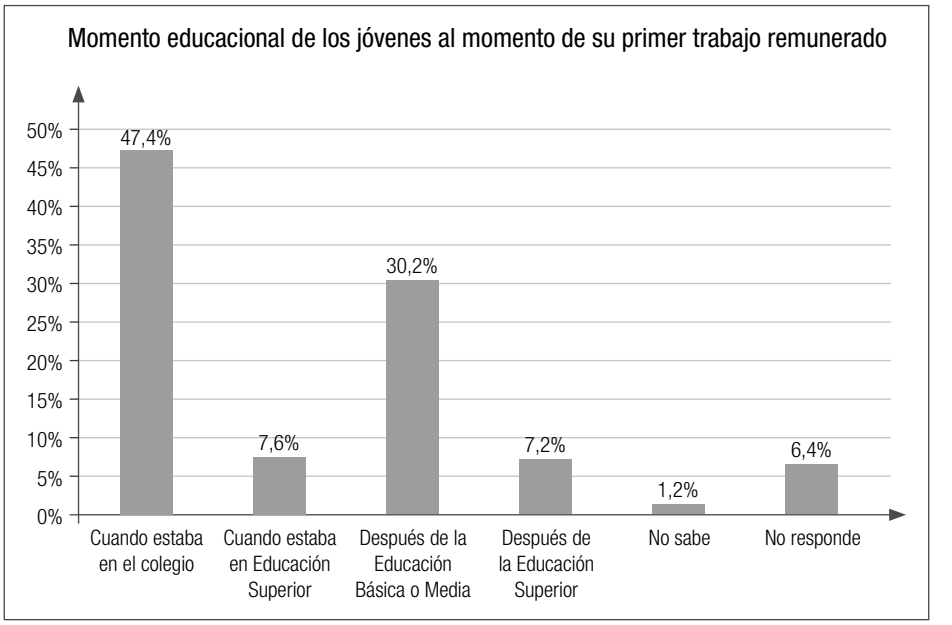
   a. Construya un gráfico de barras para estos datos. No olvide incluir todos los elementos del gráfico.
   b. Construya un gráfico circular para estos datos. No olvide incluir todos los elementos del gráfico.
   c. Considerando el tipo de datos estudiado, ¿qué tipo de gráfico, de barras o circular, cree usted que es el más adecuado?

2. Considere el siguiente gráfico que muestra los factores que los jóvenes creen que definen una Educación Básica y Media de calidad.



¿Podría representar los datos a través de un gráfico circular? Explique.

3. Considere el gráfico de barras que muestra el momento educacional de los jóvenes cuando tuvieron su primer trabajo remunerado.



Note que hay cierta ambigüedad en la categoría "Después de la Educación Básica o Media". Dada la suma de las frecuencias porcentuales que se muestran en las barras, podemos deducir que cada entrevistado debe pertenecer solo a una de las categorías. De aquí se entiende que la categoría que mencionamos significa "después de terminar la Educación Básica, en caso de que no haya asistido a Educación Media, o después de terminar la Educación Media, si es que asitió". De acuerdo a esto, responda.

a. ¿Favorece el ordenamiento de las categorías en el eje horizontal la extracción de conclusiones?

b. Replique el gráfico de barras utilizando un ordenamiento creciente de educación. Extraiga dos conclusiones de carácter general a partir de su gráfico.

c. ¿Podría representar los datos a través de un gráfico circular? Si es así, constrúyalo y luego interprete.

3.5. Diagramas de tallo y hojas

Los gráficos que hemos estudiado hasta ahora son adecuados para representar datos cualitativos. En esta sección y las siguientes, nos referiremos a representaciones para datos cuantitativos.

El *diagrama de tallo y hojas* permite la organización de conjuntos de datos cuando el número de observaciones no es muy grande, y es la única representación gráfica de las que estudiaremos que muestra los datos originales.

Presentaremos este gráfico a través de un ejemplo. Retomemos los datos sobre el número de visitas diarias de cierto sitio de Internet, observados durante un mes.

16	27	26	5	11	33	23	17	15	20
3	14	29	21	23	31	16	8	14	28
19	20	24	35	7	12	22	27	18	20

La **Figura III.21** muestra un diagrama de tallo y hojas que representa estos datos. Los números a la izquierda de la línea vertical se denominan *tallos* y los números a la derecha, *hojas*. A continuación, describiremos la construcción de este diagrama.

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

2 | 6 = visitas

Figura III.21: Diagrama de tallo y hojas del número de visitas diarias a cierto sitio de Internet.

Se comienza por hacer una lista de los tallos. En general, esta lista consiste en el o los primeros dígitos, de izquierda a derecha de los valores de las observaciones. En el ejemplo, ya que las cantidades de visitas a la página van desde **3** a **35**, los tallos consisten en los números **0**, **1**, **2** y **3**, que representan las clases de **0** a **9**, **10** a **19**, **20** a **29** y **30** a **39**, respectivamente. Anotamos estos dígitos en una columna, como se muestra en la **Figura III.22**.

0
1
2
3

Figura III.22: Se disponen los tallos de manera vertical.

Hay que asegurarse de incluir como tallos todos los valores entre el máximo y el mínimo, aun cuando no haya observaciones en dichas categorías. A modo de ejemplo, en el conjunto de datos que seguimos, si no hubiese observaciones entre **10** y **19** visitas diarias, aun incluiríamos el número **1** como parte del conjunto de tallos. Esta observación es importante, ya que al omitir una categoría en la figura, se pierde la percepción visual de la dispersión de los valores de los datos. Esto ocurre cuando no se respeta la escala en la construcción del gráfico.

Una vez representados los tallos en una columna, se deben agregar las hojas a la figura. En este caso, para cada observación en el conjunto de datos, anotamos el dígito que corresponde a las unidades, en la fila que corresponde al digito en las decenas. Por ejemplo, para el primer valor, **16**, anotamos el **6** en la fila que contiene al tallo con el número **1**, como se muestra en la **Figura III.23**.

0	
1	6
2	
3	

Figura III.23: El dato 16 se anota escribiendo las unidades, 6, a la derecha de las decenas, 1.

Los siguientes valores se anotarán de la misma manera. Notamos que, por ejemplo, el valor **5** tiene **0** decenas y **5** unidades, luego, se anota el **5** en la fila que contiene al tallo con el número **0**. La **Figura III.24** muestra un paso intermedio en la construcción, donde se ha registrado las primeras cinco observaciones: **16**, **27**, **26**, **5** y **11**.

0	5
1	6 1
2	7 6
3	

Figura III.24: Representación de los primeros cinco datos.

Las hojas deben estar alineadas verticalmente para que la longitud relativa de las filas sea percibida visualmente. De este modo, la longitud de cada fila representa la frecuencia de su clase.

Se prosigue de la misma manera con el resto de las observaciones y se obtiene la **Figura III.25**.

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

Figura III.25: Representación de todos los datos en el diagrama.

Finalmente, se ordenan las hojas de menor a mayor, se agrega la leyenda que ayuda a la lectura y el título del gráfico, como se muestra en la **Figura III.26**. La figura muestra, por ejemplo, que los datos se encuentran concentrados en las clases ente **10** y **29** visitas diarias, y que hay solo **3** días en que el sitio fue visitado, al menos, **30** veces.

Número de visitas diarias a cierto sitio de Internet

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

2 | 6 = visitas

Figura III.26: Diagrama de tallo y hojas del número de visitas diarias a un sitio de Internet.

Como mencionamos anteriormente, el diagrama de tallo y hojas presenta una característica distintiva como representación gráfica. Por una parte, nos entrega una apreciación visual de la distribución de los datos, pero, además, nos permite regresar a cada uno de sus valores. En la **Figura III.26** vemos, por ejemplo, que **3** días el sitio fue visitado **20** veces, dado que leemos **3** ceros en la fila asociada al tallo con el número **2**.

Ejercicios

1. Los siguientes datos corresponden a las edades de una muestra de **50** personas jubiladas entrevistadas durante el mes de noviembre de 2011.

71	65	66	61	54	93	60	86	70	70
73	73	55	63	56	62	76	54	82	79
76	68	53	58	80	85	56	61	61	64
65	62	90	69	76	79	77	54	64	74
65	65	61	56	63	80	56	71	79	84

   a. Construya un diagrama de tallo y hojas para estos datos. No olvide el título de la figura.
   b. ¿En qué categoría de edades se encuentra la mayor cantidad de personas jubiladas?
   c. De manera general, ¿cómo se encuentran distribuidas las edades de las personas jubiladas?

2. Las estaturas en cm de los **28** alumnos de un octavo básico corresponden a:

154	158	162	148	163	153	159	180	165	168
156	148	162	157	153	158	147	165	166	175
172	167	160	155	147	156	161	159		

   a. Construya un diagrama de tallo y hojas para estos datos. Describa la forma global de la figura.
   b. A partir de la figura, ¿dónde se concentra la mayor cantidad de observaciones?
   c. ¿Diría usted que hay una gran variación entre las estaturas de los niños del curso? Justifique su respuesta.

3.6. Diagramas de puntos e histogramas

3.6.1 Transición a partir de un diagrama de tallo y hojas

Al construir un diagrama de tallo y hojas, se introducen, de forma informal los intervalos de valores que forman las clases, como hemos definido cuando presentamos las tablas de frecuencias para variables cuantitativas. En el ejemplo sobre el número de visitas a una página de Internet, como mencionamos, se han formado 4 categorías: entre **0** y **9**, entre **10** y **19**, entre **20** y **29**, y entre **30** y **39**, y se agruparon los datos en la categoría correspondiente. De este modo, la forma del diagrama de tallo y hojas en la **Figura III.26** tiene forma similar a la que se muestra en la **Figura III.27**. Este gráfico se denomina *diagrama de puntos*.

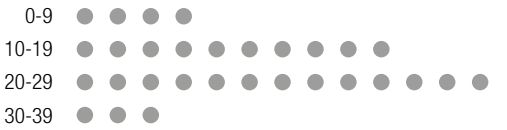


Figura III.27: Diagrama de puntos del número de visitas diarias a cierto sitio de Internet.

Reorientando el diagrama de puntos, de modo de representar las categorías a lo largo del eje horizontal, es posible relacionar esta representación con un gráfico de barras, como se muestra en la **Figura III.28**.

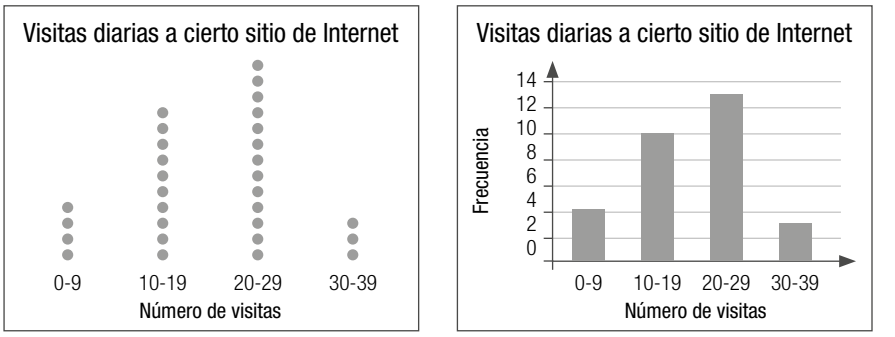


Figura III.28: Primer paso en la transición desde un gráfico de puntos a un histograma.

Cuando tratamos con variables cuantitativas, sin embargo, juntamos las barras llenando con ellas toda la escala horizontal, pues se considera a esta escala como una representación de la recta real. En este caso, dibujamos el gráfico de barras anterior como en la **Figura III.29**. Este gráfico se denomina *histograma*.

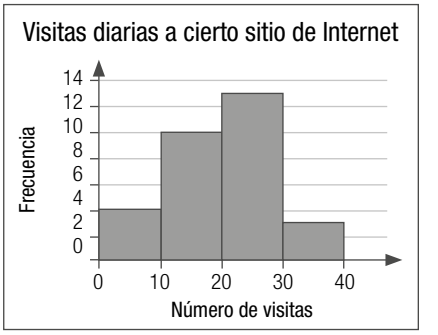


Figura III.29: Histograma del número de visitas diarias a cierto sitio de Internet.

La altura del primer rectángulo representa los días en que el número de visitas estuvo entre **0** y **10**, sin incluir el **10**; la altura del segundo rectángulo representa el número de días con visitas entre **10** y **20**, sin incluir el **20**, y así sucesivamente, imitando las categorías del diagrama de tallo y hojas. Algunos libros y programas computacionales utilizan distintas convenciones para la definición de las clases. A modo de ejemplo, la primera clase se podría definir incluyendo el valor **10** visitas, de modo que la segunda clase no lo contiene, pero sí al valor **20** visitas, y así, sucesivamente. Debido a esto, se debe tener precaución al comparar histogramas de un mismo conjunto de observaciones creados con diferentes programas computacionales.

Notemos que las clases o categorías que utilizamos al construir el diagrama de puntos que dio origen al histograma en la **Figura III.29** tienen todas la misma amplitud, de **10** visitas. Luego, las bases de las barras que forman el histograma en esta figura son del mismo ancho, también **10** visitas. Dado que la altura de cada barra corresponde a la frecuencia absoluta de la categoría (pudiendo también corresponder a la frecuencia relativa, o relativa porcentual), las áreas de las barras son proporcionales a dichas frecuencias. Esto debe cumplirse en todo histograma, de modo que la impresión visual que se tenga a partir del él sea la correcta: las categorías o clases con mayor presencia en el conjunto de datos tienen mayor representación en la figura. Retomaremos este punto en el siguiente apartado.

Por la manera en que presentamos el histograma, comenzando desde un diagrama de tallo y hojas, el ancho de las barras es de **10** unidades. Sin embargo, el ancho utilizado dependerá de cada situación particular. Al igual como discutimos cuando estudiamos tablas de frecuencias para variables cuantitativas, no existe una única manera de definir las clases o categorías a utilizar, pero debemos seguir las mismas reglas y sugerencias que se dieron en la definición de clases, en la construcción de tablas de frecuencias para este tipo de variables.

En resumen

Un conjunto de observaciones para una variable cuantitativa puede ser representado gráficamente a través de:

- *Diagrama de tallo y hojas*: permite la organización de un conjunto de observaciones no muy grande y ayuda a visualizar de forma simple la forma de la distribución.
- *Diagrama de puntos*: es similar a un diagrama de tallo y hojas, donde los valores numéricos han sido reemplazados por puntos.
- *Histograma*: es similar a un gráfico de barras para variables cualitativas. Cada barra es asociada a una de las categorías creadas. Por tratarse de variables cuantitativas, no deben existir espacios entre las barras.

3.6.2 Errores y dificultades en la construcción de histogramas

En lo que sigue, nos referiremos a algunos errores y dificultades relacionadas a la construcción e interpretación de histogramas.

- *Utilizar clases de distinto ancho o amplitud sin ajustar las alturas de las barras.* Como ya mencionamos, las áreas de las barras de un histograma deben ser proporcionales a las frecuencias (absolutas, relativas, o relativas porcentuales) de las categorías a las que ellas representan. Mostraremos la importancia de este punto retomando el ejemplo que seguimos, sobre el número de visitas a cierto sitio de Internet.

Supongamos que definimos la primera clase como "de **0** a **19** visitas", sin alterar los límites de la tercera y la cuarta clase. En este caso, existen **14** observaciones en la primera categoría, y todas las frecuencias quedan como en la **Tabla III.41**.

Número de visitas	Frecuencia
0 a 19	14
20 a 29	13
30 a 39	3

Tabla III.41: Frecuencias absolutas del número de visitas diarias a cierto sitio de Internet durante un mes, con categorías o clases redefinidas.

Si no prestamos atención a la condición de áreas proporcionales de las barras y dejamos que la altura de estas corresponda a las frecuencias en la **Tabla III.41**, el histograma asociado sería el que se muestra en la **Figura III.30**. Este histograma es erróneo, puesto que, dado que la amplitud de la primera categoría es el doble de las amplitudes de las categorías restantes, la base de la barra asociada es también el doble, lo que hace que la categoría de **0** a **19** visitas esté sobrerrepresentada en la figura. De hecho, su área es exactamente el doble de la que correspondería si nos adherimos a la condición de áreas proporcionales. Luego, para mantener la condición de áreas proporcionales, la altura del rectángulo de la categoría de **0** a **19** visitas debiese ser de $\frac{14}{2} = 7$ visitas.

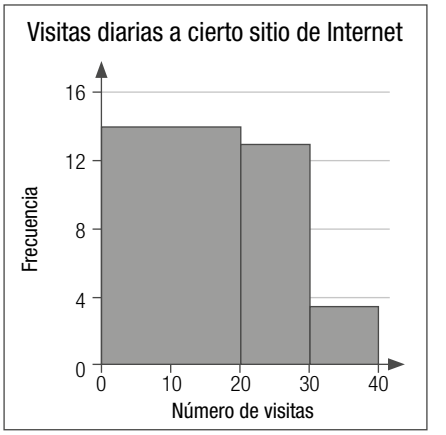


Figura III.30: Histograma erróneo del número de visitas a un sitio de Internet. La amplitud de la primera categoría es el doble de las amplitudes de las dos restantes.

Según esto, recalcamos que todas las discusiones desarrolladas en esta sección asumen clases o categorías de igual amplitud y que, salvo casos particulares, desaconsejamos utilizar categorías o barras de diferentes amplitudes.

- *Construir histogramas para variables cualitativas*. El hecho de que los valores asociados a una variable se representen con números no significa necesariamente que esta sea una variable cuantitativa. Este es el caso, por ejemplo, del número de alumno o número de matrícula entregado por ciertas universidades a cada uno de ellos; el número de alumno corresponde a una variable cualitativa, sin embargo, un error es representar su distribución a través de un histograma, como se muestra en la **Figura III.31**.

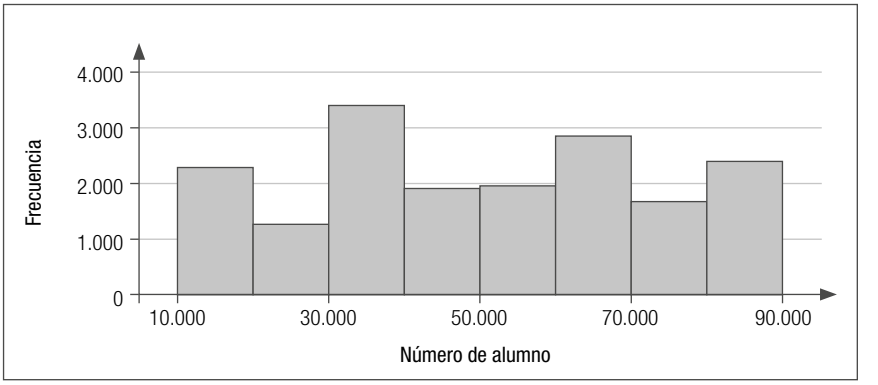


Figura III.31: Error: histograma construido para una variable cualitativa.

- *Elección del tamaño apropiado del intervalo*. Como estudiamos anteriormente, la elección del tamaño del intervalo no es única, y esto puede constituir una dificultad en su construcción. Del mismo modo, utilizar distinto número o tamaño de intervalos puede cambiar drásticamente la representación de datos. En efecto, la **Figura III.32** muestra dos histogramas para los mismos datos construidos sobre intervalos de diferentes tamaños. No existe una regla general para determinar cuál de las dos representaciones es más apropiada, sin embargo, podemos notar que probablemente el histograma de la derecha nos muestra demasiada información, o demasiada variabilidad en las alturas de las barras, cuando, en realidad, queremos tener una visión más general de la situación.

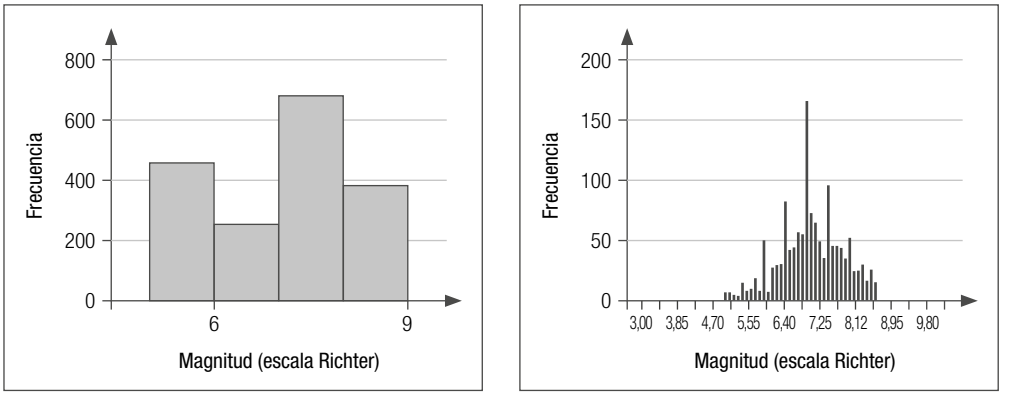


Figura III.32: Histogramas para el mismo conjunto de datos construidos en base a dos grupos de categorías o clases de diferentes amplitudes.

Ejercicios

1. Considere nuevamente los datos correspondientes a las edades de un conjunto de **50** personas jubiladas entrevistadas durante el mes de noviembre de 2011.

71	65	66	61	54	93	60	86	70	70
73	73	55	63	56	62	76	54	82	79
76	68	53	58	80	85	56	61	61	64
65	62	90	69	76	79	77	54	64	74
65	65	61	56	63	80	56	71	79	84

   a. Construya un gráfico de puntos a partir del diagrama de tallo y hojas construido anteriormente para estos datos (puede hacerlo ahora en caso de que no lo haya hecho antes) y, a partir de este diagrama de puntos, obtenga un histograma. No olvide: títulos, nombres, escalas y límites de las clases en los ejes.

   b. ¿Entre qué edades se encuentra la mayor cantidad de personas jubiladas?

   c. ¿Debió regresar al diagrama de tallo y hojas para determinar exactamente los límites de las edades en el apartado b.?

   d. ¿Qué conclusión general puede extraer a partir de un patrón observado en la figura?

2. En el problema anterior construya un histograma donde las categorías de las edades tengan una amplitud de **5** años. ¿Entre qué edades se encuentra ahora la mayor cantidad de personas? ¿es consistente con lo que observó en el problema anterior?

3. Repita el ejercicio anterior, ahora utilizando clases que vayan de **20** en **20** años.

4. Considere nuevamente la tabla de frecuencias sobre los sueldos iniciales para **50** profesionales recién graduados.

Sueldo	Frecuencia
menos de $300.000	1
$300.000 a $599.999	16
$600.000 a $899.999	20
$900.000 a $1.199.999	9
$1.200.000 a $1.499.999	4

   a. Construya un histograma en base a la frecuencia absoluta de los sueldos. No olvide nombres, graduación de los ejes y el título del gráfico.

   b. Obtenga las frecuencias relativas porcentuales y construya un histograma en base a ellas. No olvide nombres, graduación de los ejes y el título del gráfico. Utilice la misma escala que en el apartado anterior.

   c. Compare la forma de los dos histogramas obtenidos. ¿Qué puede concluir?

   d. ¿Qué porcentaje de profesionales recién graduados recibe entre $**300.000** y $**899.000**?

e. ¿A partir de cuál de los dos gráficos resulta más fácil o directo responder el apartado anterior?

f. ¿Qué categoría o clase es la más frecuente?

g. ¿Cuál(es) de los dos gráfico(s) utilizó o puede utilizar para responder el apartado anterior?

5. Los siguientes datos corresponden al número de días de la semana en que los 20 niños del tercero B practican deporte.

1	4	4	1	1	3	4	3	1	1	3
		2	11	1	2	3	3	3		

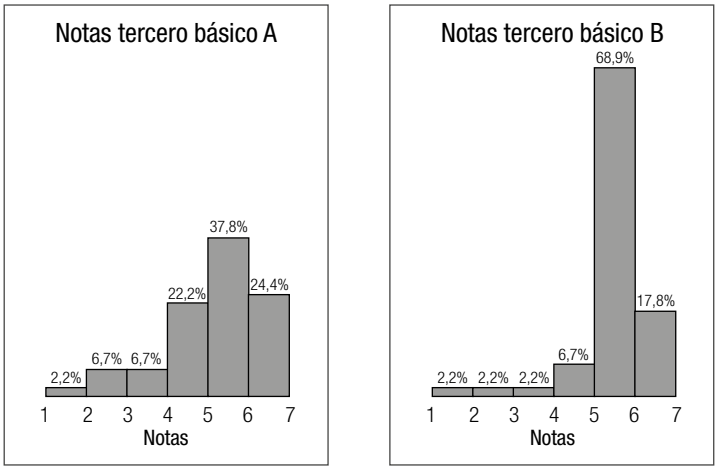
a. Si desea construir un histograma que comunique el número de niños que practican deporte 4 veces a la semana, ¿qué intervalos de clase debiese utilizar? ¿por qué?

b. Si desea que el gráfico comunique el número de niños que practican deporte 4 veces a la semana, ¿qué frecuencias utilizaría: absolutas, relativas o relativas porcentuales?

c. Si se desea que el gráfico comunique el porcentaje de niños que practican deporte 1 o 2 veces a la semana, ¿qué posible(s) intervalos puede utilizar?

d. Si se desea que el gráfico comunique el porcentaje de niños que practican deporte 1 o 2 veces a la semana, ¿qué frecuencias utilizaría: absolutas, relativas o relativas porcentuales?

6. Considere los siguientes histogramas correspondientes a las notas en una misma evaluación en la asignatura de Música, de los terceros A y B de una escuela.



a. ¿Qué porcentaje de niños del tercero A obtuvo notas entre 4 y 7? ¿y del tercero B?

b. Los alumnos que obtuvieron notas entre 6 y 7 recibirán un estímulo por su buen rendimiento. ¿Qué porcentaje del tercero A recibirá este estímulo? ¿y del tercero B?

c. Si se decide entregar este estímulo únicamente a los alumnos que obtuvieron notas entre 6,5 y 7, ¿puede determinar qué porcentajes de alumnos del tercero A y del tercero B lo recibirán? Explique.

3.7. Gráficos de líneas o de tendencia

Las representaciones gráficas que hemos estudiado hasta ahora tienen en común que su objetivo es representar las frecuencias de ocurrencia de los valores de las observaciones. De este modo, las figuras estudiadas corresponden a una manera de representar la distribución de las observaciones en el conjunto de datos.

Los *gráficos de líneas* o *de tendencia* que aquí estudiaremos tienen un objetivo diferente. Este tipo de gráfico se utiliza para estudiar la evolución de las observaciones, usualmente, a través del tiempo.

A modo de ejemplo, consideremos la **Figura III.33**, que corresponde a un gráfico de líneas del Índice de Costos del Transporte en Chile entre los años 2010 y 2011[17]. El eje horizontal, o de las abscisas, representa el instante en que fueron tomadas las observaciones, mientras que el eje vertical o de las ordenadas representa el valor de las observaciones mismas. Si bien los datos son obtenidos en instantes discretos de tiempo, las observaciones se unen a través de líneas, para ayudar al lector a percibir la tendencia general del comportamiento de la variable.

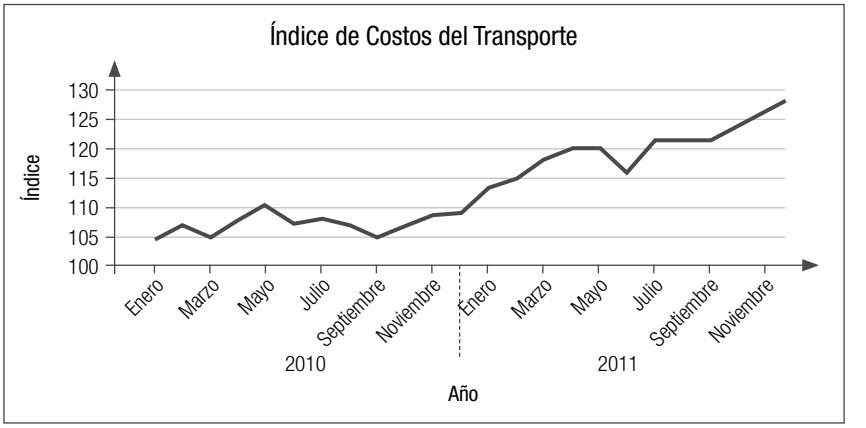


Figura III.33: Gráfico de líneas o tendencia del Índice de Costos del Transporte en Chile entre los años 2010 y 2011.

En la figura, la línea muestra la evolución del Índice de Costos del Transporte a través del tiempo. A modo de ejemplo, del gráfico leemos que en enero del año 2010, este índice correspondía aproximadamente a **105**, o que en mayo de 2011, correspondía a aproximadamente **120**. Aunque el índice crece y decrece en distintos intervalos de tiempo, la figura muestra una tendencia global creciente.

Se debe prestar especial atención a posibles distorsiones de la información entregada por este tipo de gráficos, que se deben a las escalas utilizadas en los ejes. A modo de ejemplo, los mismos datos sobre el Índice de Costos del Transporte 2010 y 2011 son representados en la **Figura III.34** utilizando una ampliación vertical. En este segundo gráfico, se da la idea de un crecimiento del índice mucho más acelerado.

[17] El Índice de Costos del Transporte en nuestro país cuantifica los precios de una canasta de bienes y servicios fija consumida por el sector transporte de carga por carretera dentro de las fronteras del país. Los datos fueron tomados del sitio de Internet del Instituto Nacional de Estadísticas.

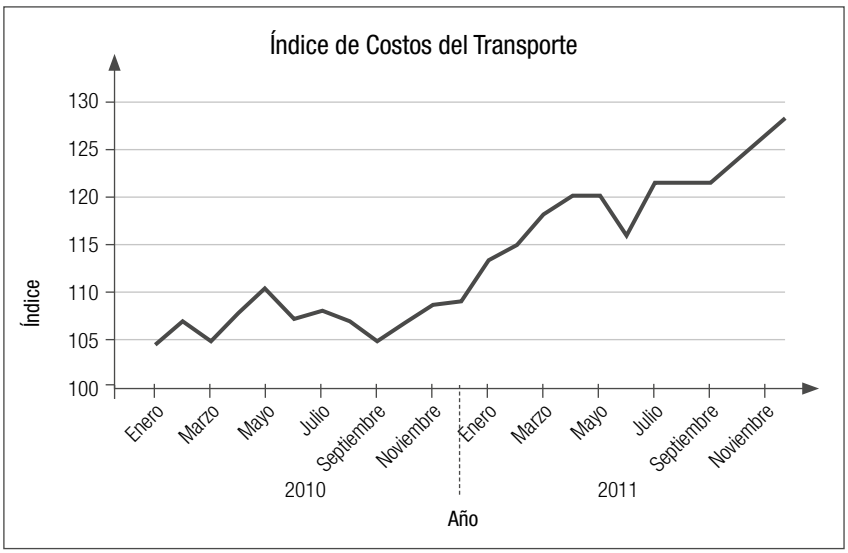


Figura III.34: Gráfico de líneas o tendencia del Índice de Costos del Transporte en Chile entre los años 2010 y 2011. La separación de las unidades en la escala vertical es mayor que en la Figura III.33, lo que sugiere un crecimiento más acelerado.

Si bien no es posible decidir cuál de las dos **Figuras, III.33** o **III.34**, es la correcta, sí podemos afirmar que en caso de querer comparar el comportamiento de dos series de valores, es preferible presentar ambas series en el mismo gráfico. A modo de ejemplo, la **Figura III.35** muestra el comportamiento del Índice de Precios al Consumidor, IPC, para los años 2010, 2011 y 2012. Representar los 3 años en la misma figura permite comparar los crecimientos en años diferentes. En este caso, se observa, por ejemplo, que durante el primer semestre del año 2012, el IPC creció más lento que en el mismo período de los años 2010 y 2011, dado que la serie de 2012 tiene menor pendiente que las restantes en el período de enero a junio. También observamos que, en general, el IPC creció más lento durante 2010 que durante 2011, dado que la primera de estas series tiene una pendiente menos pronunciada.

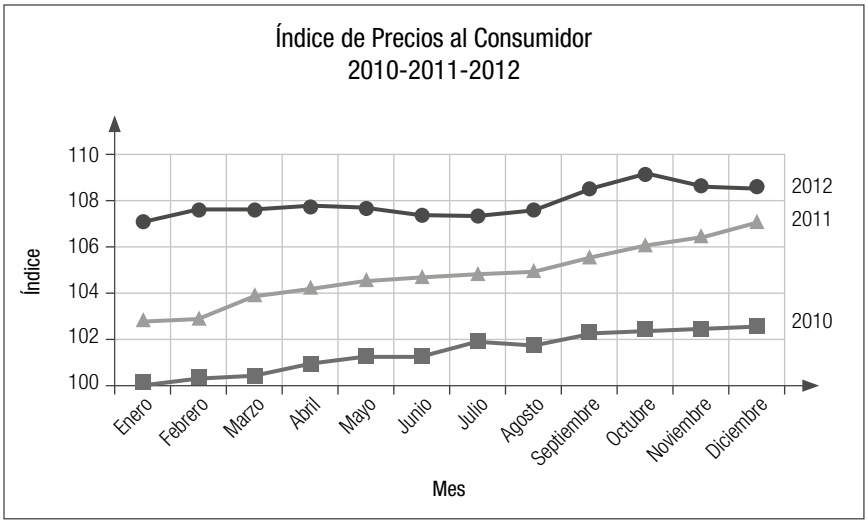


Figura III.35: Índice de Precios al Consumidor, años 2010 a 2012.

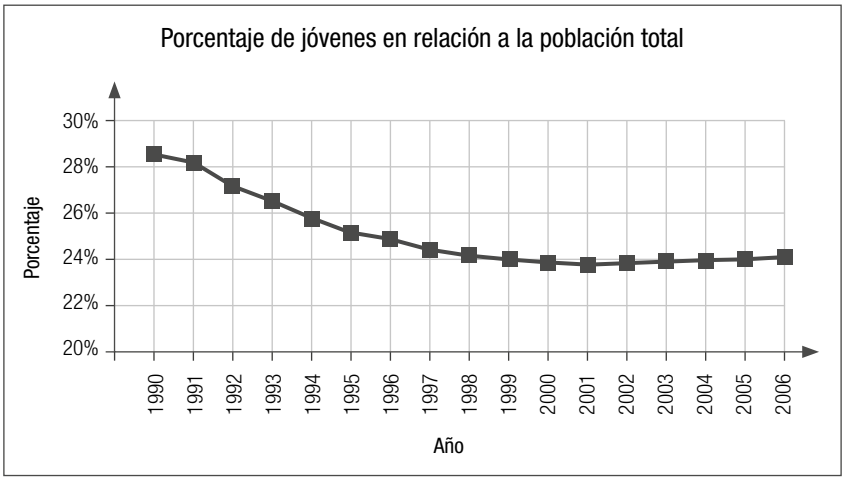
En resumen

- Un *gráfico de líneas* o *de tendencia* corresponde a una representación gráfica de variables cuantitativas, donde los valores han sido tomados en algún orden que, generalmente, corresponde al tiempo.
- Los valores son unidos a través de líneas rectas, lo que facilita la identificación de tendencias o patrones.
- Los gráficos de líneas se diferencian de los gráficos estudiados anteriormente en que su objetivo no es representar la distribución de las observaciones, sino mostrar la evolución de estas a través del tiempo.
- Para comparar el comportamiento de dos series, se sugiere representarlas en la misma figura y así evitar conclusiones erróneas debido a diferencias de escala.

Ejercicios

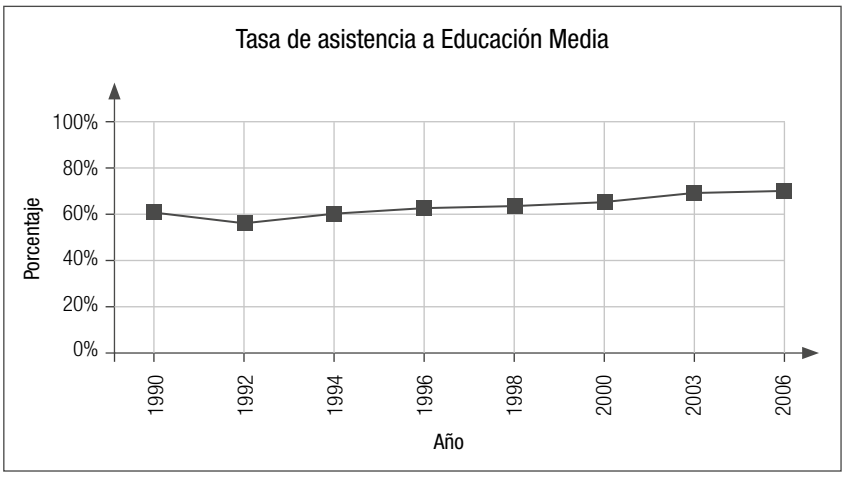
En lo que sigue, analizaremos representaciones gráficas presentadas en el informe de la "5ta encuesta nacional de la juventud" del año 2006. Recordemos que las edades de los jóvenes considerados varían entre los **15** y **19** años.

1. Considere la figura sobre la evolución del tamaño de la población joven del país (medida como porcentaje del total de la población) entre los años 1990 y 2006.

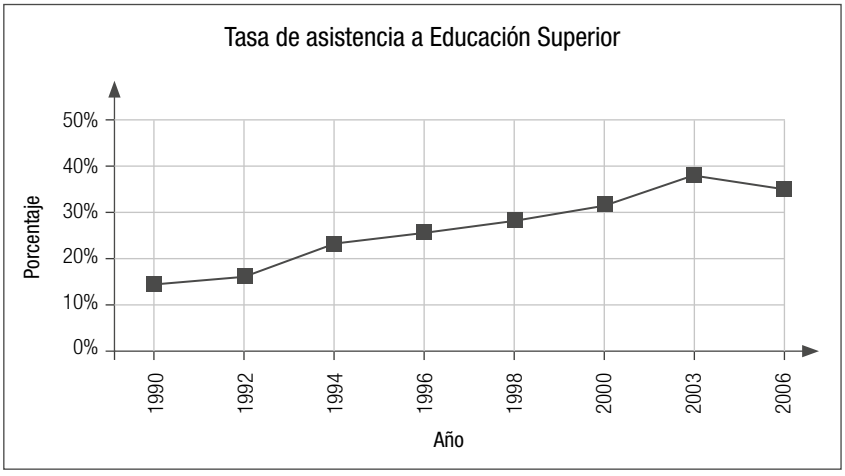


   a. ¿Cuál ha sido la evolución del porcentaje de jóvenes en Chile entre los años 1990 y 2000?
   b. ¿Cuál ha sido la evolución del porcentaje de jóvenes en Chile entre los años 2000 y 2007?
   c. ¿Puede aventurar explicaciones para este aparente cambio en la tendencia?

2. Considere la siguiente figura, que muestra la evolución de las tasas netas de asistencia a Educación Media entre los años 1990 y 2006[18]. Comente esta evolución.



3. Considere la siguiente figura, que muestra la evolución de las tasas netas de asistencia a Educación Superior entre los años 1990 y 2006[19]. Comente esta evolución.



4. ¿Puede comparar las evoluciones de las tasas de asistencia a Educación Media y Superior directamente a partir de los gráficos que se muestran? (ponga atención a los ejes).

5. De acuerdo a los datos que se leen en los gráficos de tendencia de las tasas de asistencia a Educación Media y Superior, construya un único gráfico de tendencias que muestre a ambas. Extraiga al menos 2 comparaciones a partir de su gráfico.

[18] La tasa neta es calculada sobre todas las personas en edad de asistir.

[19] La tasa neta es calculada sobre todas las personas en edad de asistir.

3.8. Gráficos de dispersión

Cuando estudiamos tablas de frecuencias, vimos que las tablas de doble entrada son de utilidad para estudiar la relación o posible asociación, entre dos variables cualitativas. En el caso de las variables cuantitativas, es posible estudiar la relación entre dos variables de manera gráfica, a través de un *gráfico de dispersión*. De este modo, al igual como lo señalamos al presentar los gráficos de línea o tendencia, los gráficos de dispersión son de distinta naturaleza que las representaciones gráficas estudiadas con anterioridad, dado que su objetivo no corresponde a la visualización de la distribución de las observaciones.

Supongamos que estamos interesados en estudiar la relación entre los resultados en las pruebas SIMCE[20] de Lenguaje y Matemática de cuartos básicos, para una misma región en el año 2011. Para ello, disponemos, por ejemplo, de los porcentajes de alumnos en cada región que obtuvieron puntajes en la categoría denominada Avanzada. La **Tabla III.42** muestra los porcentajes en la categoría Avanzada en cada una de las dos pruebas, para las diferentes regiones.

Región	Lenguaje (%)	Matemática (%)
XV	43	32
I	38	26
II	40	27
III	36	23
IV	42	28
V	40	27
VI	40	29
VII	43	32
VIII	44	32
IX	42	27
XIV	42	26
X	43	30
XI	41	26
XII	40	29
RM	44	32

Tabla III.42: Porcentajes de alumnos con puntajes en categoría Avanzada en las pruebas SIMCE de Lenguaje y de Matemática, año 2011, por región.

En el tipo de problemas que seguimos, una observación está constituida por dos valores, uno de cada una de las variables de interés. En nuestro caso, una observación corresponde a un par formado por el porcentaje de alumnos en categoría Avanzada en la prueba de Lenguaje, y el porcentaje de alumnos en categoría Avanzada en la prueba de Matemática, ambos para la misma región. A modo de ejemplo, en la **Tabla III.42** leemos que la observación asociada a la III Región corresponde a los valores **36%** y **23%**. Notamos que existe una observación asociada a cada una de las regiones, es decir, se tienen **15** observaciones.

[20] El SIMCE, o sistema de Medición de los Resultados de Aprendizaje, fue creado en 1988 con el objetivo de intitucionalizar diversas iniciativas en el ámbito de la evaluación escolar en Chile.

La **Figura III.36** muestra la representación gráfica de estas observaciones a través de un gráfico de dispersión.

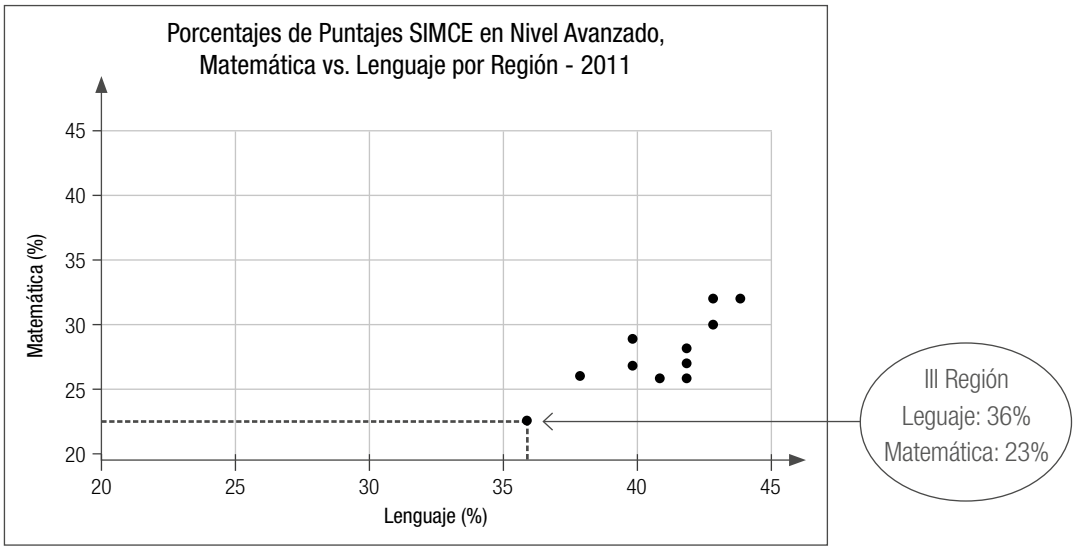


Figura III.36: Gráfico de dispersión del porcentaje de alumnos en la categoría Avanzada en prueba SIMCE de Matemática versus el porcentaje de alumnos en la categoría Avanzada en la prueba SIMCE de Lenguaje, ambos en el año 2011.

En un gráfico de dispersión, cada una de las variables es asociada a un eje. En el ejemplo, el porcentaje en nivel Avanzado en Lenguaje ha sido asociado al eje horizontal, o de las abscisas, mientras que el porcentaje en nivel Avanzado en Matemática ha sido asociado al eje vertical, o de las ordenadas.

En un gráfico de dispersión, ubicamos cada una de las observaciones en el plano cartesiano creado, y las marcamos con un punto o marca pequeña. A modo de ejemplo, en la **Figura III.35** hemos indicado la observación asociada a la III Región que, como dijimos, corresponde al par de valores **36%** y **23%**. El gráfico debe incluir un título y el nombre, unidad de medida y valores, de cada uno de los ejes.

En la **Figura III.36**, vemos que existe una relación creciente entre los porcentajes pertenecientes al nivel Avanzado en Lenguaje y Matemática. Esto es, a mayor porcentaje de alumnos en nivel Avanzado de Lenguaje, mayor tiende a ser el porcentaje de alumnos en nivel Avanzado de Matemática en la misma región, y viceversa.

En general, es posible identificar variables que tengan una relación creciente, como la de la **Figura III.36**, decreciente, combinada o nula. La **Figura III.37** muestra datos que representan el peso de **10** hijos versus el peso de sus padres. A modo de ejemplo, se identifica una observación que representa al peso de un padre (**60 kg**) y el peso de su hijo (**63,6 kg**). La figura muestra un patrón creciente, lo que significa que, en general, mientras mayor es el peso del padre, mayor es también el peso del hijo.

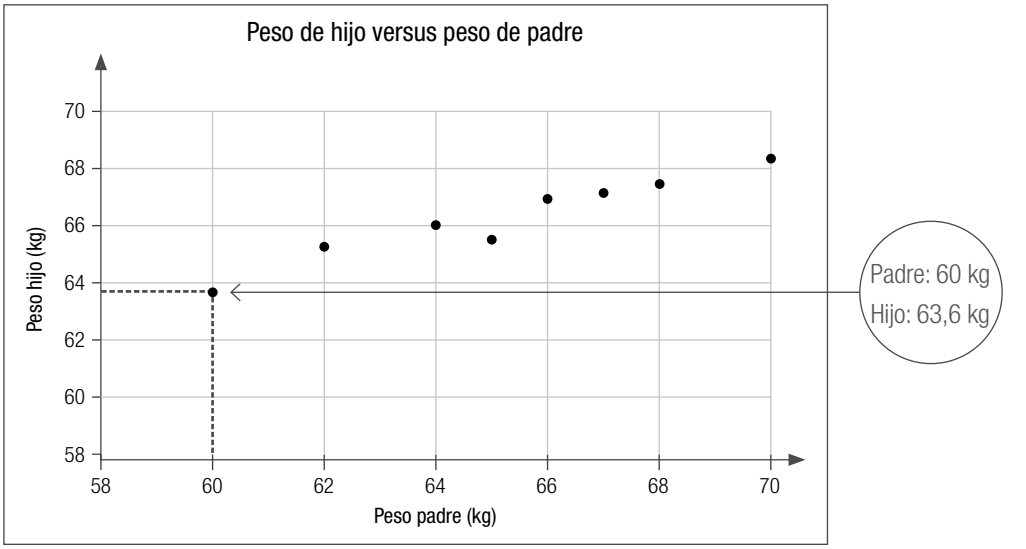


Figura III.37: Peso de 10 hijos versus el peso de sus respectivos padres (kg).

La **Figura III.38** muestra la relación entre la velocidad que alcanzan los automóviles en cierta avenida y la densidad del tráfico en la misma. El gráfico muestra un patrón decreciente, lo que significa que mientras mayor sea la densidad del tráfico, es decir, mientras mayor sea la congestión vehicular, menor es, en general, la velocidad a la que circulan los automóviles.

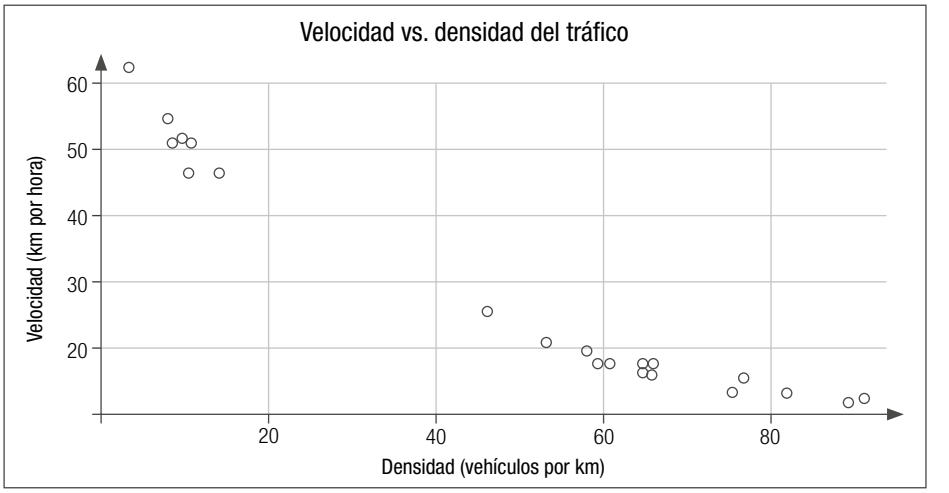


Figura III.38: Velocidad alcanzada por los automóviles, versus densidad del tráfico, en cierta avenida.

En general, cuando es posible identificar un patrón en el gráfico, decimos que las variables en cuestión están asociadas. Cuando no existe una asociación entre las variables en estudio, su gráfico de dispersión corresponde a una nube y se visualiza como en la **Figura III.39**, que muestra la relación entre las notas obtenidas por los alumnos de un curso en dos evaluaciones, cada una en una escala de 1 a 7.

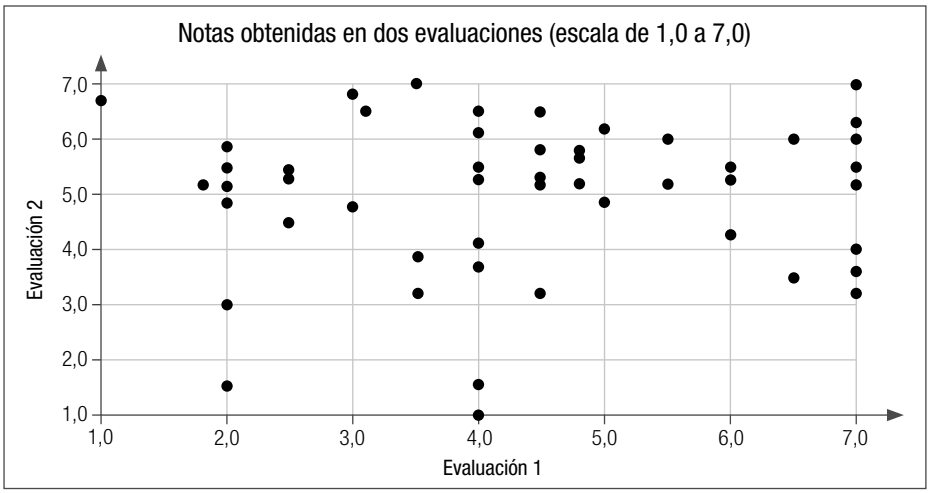


Figura III.39: Notas obtenidas por los alumnos en dos evaluaciones de un ramo. La figura muestra que no existe asociación entre las notas de estas evaluaciones. Cada punto representa las notas de un mismo alumno.

La asociación entre dos variables es un concepto simétrico, y esto hace que se puedan intercambiar las variables con los ejes, representando cualquiera de las dos variables en el eje de las abscisas y lo mismo con el eje de las ordenadas. Sin embargo, existen situaciones en las que se puede pensar que una de las variables es *causa* de la otra.

A modo de ejemplo, supongamos que estamos interesados en estudiar la relación entre el hábito de fumar (medido en número de cajetillas al día, per cápita) y la tasa de cáncer, por país. En esta situación, en caso de existir una asociación, lo razonable es pensar que la intensidad del hábito de fumar es causa de las tasas de cáncer observadas. En estos casos, se acostumbra a graficar la variable *causa* en el eje de las abscisas, y la variable *efecto* en el eje de las ordenadas.

Notemos que hemos usado las palabras *causa* y *efecto* en letra cursiva. Esto se debe a que determinar estadísticamente la causalidad de una variable sobre otra es bastante complejo y solo se logra a través de experimentos especialmente diseñados con este fin. Cuando no es ese el caso, solo decimos que las variables están asociadas, y no podemos adjudicar causalidad ni a una ni a la otra.

[21] Adaptado de http://humanidades.cchs.csic.es/cchs/web_UAE/errorcomun/errorcomun.htm

Consideremos, a modo de ejemplo, los datos en la **Figura III.40**, que muestran el número de nacimientos anuales y el número de nidos de cigüeñas encontrados en el mismo período, para diferentes localidades[21]. El gráfico muestra un marcado patrón creciente entre ambas variables, es decir, en general, mientras mayor es el número de nidos de cigüeñas en una localidad, mayor es el número de nacimientos observado durante dicho año en la misma. Decimos, entonces, que el número de nacimientos anuales y el número de nidos encontrados durante el mismo período anual en una misma localidad se encuentran *asociados*.

Ciertamente, esto no significa que un mayor número de nidos de cigüeña es causa de un mayor número de nacimientos dentro de una misma localidad. En efecto, es posible que la presencia de la asociación detectada se deba a la extensión de las localidades (a que localidades más extensas presenten simultáneamente un mayor número de nacimientos y de nidos de cigüeñas), a la ruralidad de las localidades (a que localidades de carácter más rural posean simultáneamente un mayor número de nacimientos y de nidos de cigüeñas), entre otras posibles causas. Luego, de la **Figura III.40**, podemos concluir que el número de nacimientos anuales y el número de nidos de cigüeñas observados durante un año en una localidad están asociados, pero no podemos concluir que uno de ellos sea la causa del otro.

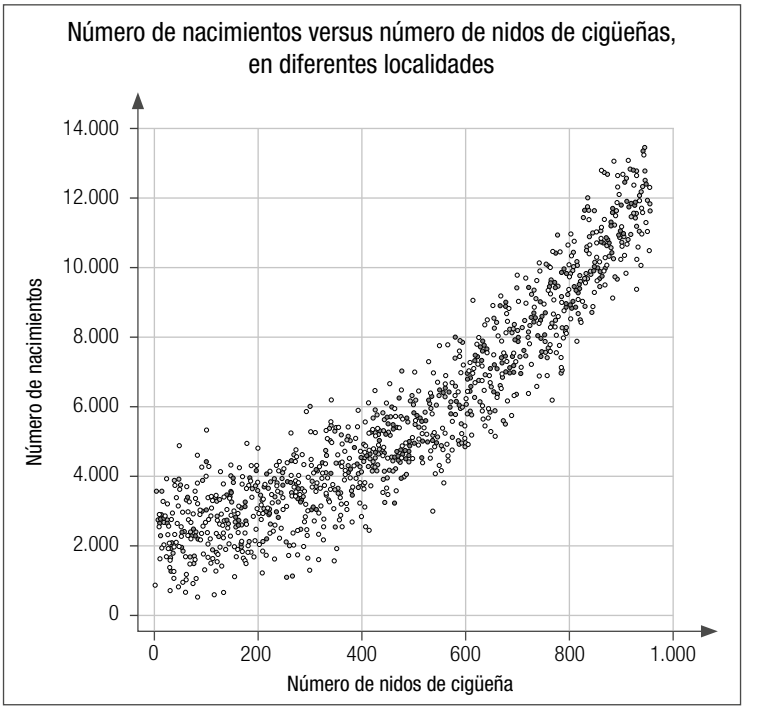


Figura III.40: Número de nacimientos anuales versus el número de nidos de cigüeñas observados en el mismo período, en diferentes localidades.

En resumen

- Un *gráfico de dispersión* muestra la relación entre dos variables cuantitativas.
- Cada eje representa una de las variables y cada par de datos es representado a través de un punto en el plano cartesiano.
- La presencia de un patrón en la figura muestra la existencia de *asociación* entre las variables, sin embargo, *no es evidencia de que una sea la causa de la otra.*

Ejercicios

1. Considere las siguientes situaciones y discuta el tipo de gráfico de dispersión que esperaría: asociación creciente o decreciente, o falta de asociación.
   a. Se obtienen observaciones sobre las tasas de cáncer en diferentes países y los niveles medios de consumo de cigarrillos de su población.
   b. Se obtienen observaciones sobre la cantidad de agua caída en cierta localidad y el flujo del río que la cruza.
   c. Se obtienen observaciones sobre número de horas de estudio y el número de horas de juego en computador de cada niño de un curso.
   d. Se obtienen observaciones sobre las temperaturas diarias medias en una ciudad y su tasa de accidentes automovilísticos.
   e. Se obtienen observaciones sobre el número de horas dedicadas por cada alumno de un curso al estudio de matemática y su rendimiento obtenido en esta asignatura.
   f. Se obtienen observaciones sobre la temperatura corporal y la presión arterial de un grupo de personas.
   g. Se obtienen observaciones sobre el número de años de escolaridad de un grupo de apoderados y su altura.
   h. Se obtienen observaciones sobre el número de horas dedicadas al entrenamiento de un grupo de atletas y sus tiempos registrados en la prueba de 400 metros planos.
2. Proponga un par de variables que usted crea que tienen una relación creciente entre sus valores.
3. Proponga un par de variables que usted crea que tienen una relación creciente entre sus valores.

3.9. Consideraciones generales sobre representaciones gráficas

Es importante considerar que cualquiera sea la representación gráfica utilizada, esta debe ser integral, en el sentido de que "no debe mentir" ni distorsionar la realidad, y debe ser construida de modo que favorezca la correcta interpretación y contribuya a responder la pregunta de interés.

De este modo, las escalas utilizadas en una representación gráfica deben ser consistentes; es decir, deben variar según intervalos constantes. Las escalas de gráficos que deben ser comparados entre sí, deben ser iguales. Además se debe prestar especial atención a las escalas de los gráficos entregados por programas computacionales no estadísticos. Se debe tener presente, también, que los símbolos utilizados deben facilitar la lectura. Esto se conecta con la capacidad de comprender y relacionar elementos de la cultura del receptor. Del mismo modo, una representación gráfica debe mantener la simplicidad, y solo se deben incluir los elementos que aportan información relevante y ayudan a la interpretación.

Debido a lo anterior, muchos autores e investigadores han tratado de definir los elementos necesarios para la representación de variables. Así, por ejemplo, se puede señalar que un gráfico queda determinado por los siguientes elementos:

- Las palabras que aparecen en el gráfico, como su título, las etiquetas de los ejes y de las escalas, que proporcionan las claves necesarias para comprender las relaciones representadas.
- El contenido matemático subyacente. Por ejemplo, los conjuntos numéricos empleados, las escalas, las unidades de medida y otros conceptos matemáticos que el estudiante necesita dominar para interpretarlo.
- Los convenios específicos que se usan en cada tipo de gráfico y que se deben conocer para poder realizar una lectura o construcción correcta. Por ejemplo, el alumno necesita saber cuántas unidades representa una unidad icónica o qué representa cada categoría.

Ejercicio

Para las siguientes situaciones y conjuntos de observaciones, sugiera qué tipo de representación gráfica utilizaría y bosquéjela.

a. Se desea estudiar la evolución de las ventas en un supermercado. Para ello, se observa sus ventas diarias, desde el 1 de enero hasta el 31 de diciembre de un mismo año.
b. Se quiere analizar si existe una relación entre el sexo de los alumnos y la carrera a la que ingresan. Para ello, se consideran 3 carreras diferentes y se registra el sexo de cada uno de sus alumnos.
c. Se observa los porcentajes de votos recibidos por cada uno de los 3 candidatos a la alcaldía de cierta comuna.
d. Se desea saber si existe una relación entre el peso y la altura de los niños en la selección de básquetbol. Para ello, se observa la altura y peso de cada uno de estos deportistas.

e. Se quiere representar gráficamente el tipo de programa favorito de un conjunto de personas seleccionadas de manera aleatoria, donde cada una debe elegir entre series, caricaturas, musicales y noticieros.

f. Se desea estudiar el número de satélites de cada uno de los planetas del sistema solar.

g. Se entrevista a un grupo de trabajadores en una empresa y se les pide que se identifiquen con una de tres categorías dependiendo de su nivel de actividad física: no practicante, medianamente practicante y practicante. Se desea estudiar si esta característica se encuentra asociada con la edad de los trabajadores, por lo que se les pide que indiquen en qué tramo de edad se encuentran: menor de 20 años, entre 20 y 29 años, entre 30 y 39 años, 40 años o más.

3.10. Lectura de gráficos

Recordemos que uno de los objetivos de las representaciones gráficas y, probablemente, el más importante corresponde a comunicar información. Una representación de este tipo debe constituir una fuente que habilite a un receptor, para responder ciertas preguntas relevantes. En este sentido, el receptor debe ser capaz de leer una representación gráfica.

La literatura actual indica que existen, principalmente, tres niveles de lectura de gráficos:

- *Nivel Elemental*: lectura directa e inmediata del gráfico, sin interpretación de la información que contiene.
- *Nivel Intermedio*: lectura que requiere interpretación e integración de los datos que muestra el gráfico.
- *Nivel Avanzado*: lectura que requiere inferencias que van más allá de los datos contenidos en el gráfico.

Para ejemplificar cada uno de estos niveles de lectura, utilizaremos el gráfico de barras en la **Figura III.41**, que muestra la distribución de los medios de transporte que utilizan los niños de un cuarto básico para llegar a la escuela durante el mes de noviembre.

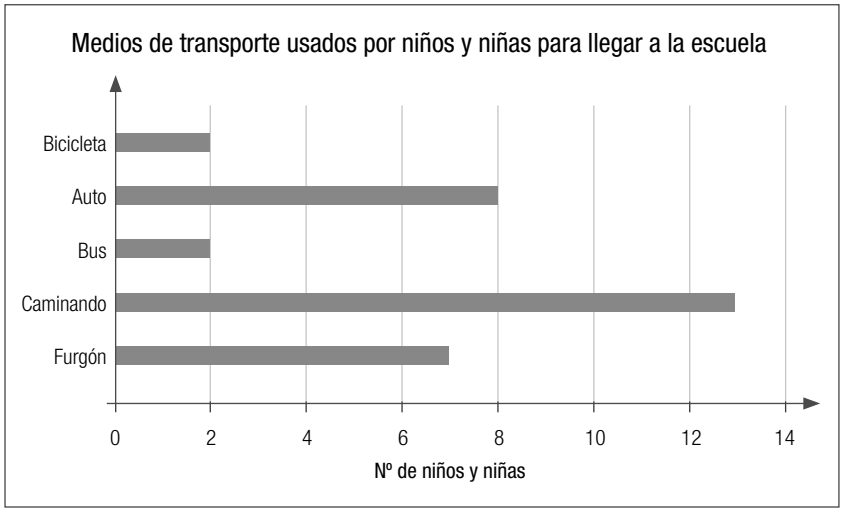


Figura III.41: Medios de transporte utilizados por los niños del cuarto básico para llegar a la escuela durante el mes de noviembre.

La **Tabla III.43** presenta ejemplos de preguntas de cada uno de los niveles de lectura descritos anteriormente, Elemental, Intermedio y Avanzado, que se pueden responder a partir de la **Figura III.41**.

Nivel de lectura	Pregunta	Respuesta
Elemental	¿Cuántos niños del cuarto básico van en bus a la escuela?	2 niños
Intermedio	¿Cuáles son los dos medios de transporte más utilizados para llegar a la escuela por los niños del cuarto básico?	Auto y caminando.
Avanzado	¿Cree usted que el mismo gráfico sufriría cambios si se representaran los medios de transporte utilizados por los niños durante el mes de julio, en lugar de noviembre? Justifique.	Es posible, por ejemplo, aventurar que el número de niños que se va a la escuela caminando o en bicicleta disminuirá en julio por ser un mes de clima más frío y húmedo. En particular, es probable que el medio preferido deje se ser la caminata.

Tabla III.43: Algunas preguntas que pueden responderse a partir del gráfico de la Figura III.41, y los niveles de lectura a los que corresponden.

En otro caso, retomemos el ejemplo de la **Figura III.35**, que reproducimos aquí como **Figura III.42**, que muestra la evolución del Índice de Precios al Consumidor, IPC, a través de los años 2010, 2011 y 2012.

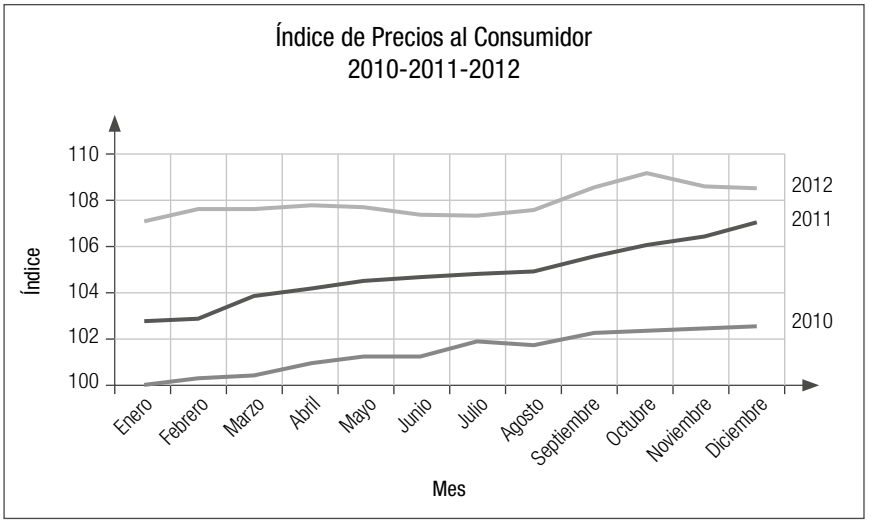


Figura III.42: Índice de Precios al Consumidor, años 2010-2012.

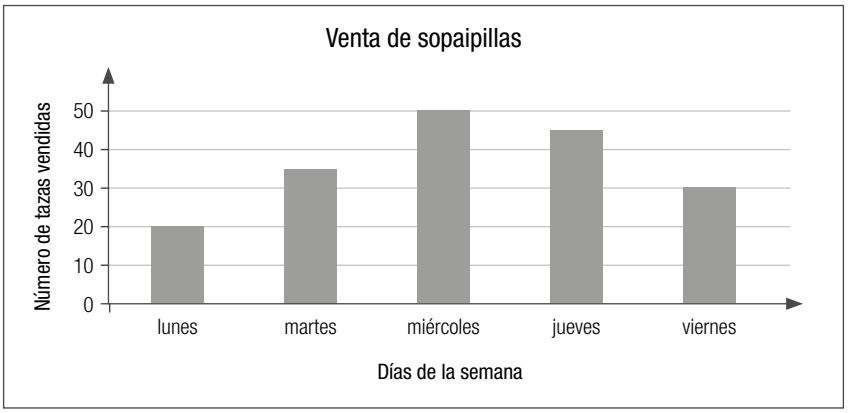
La **Tabla III.44** presenta ejemplos de preguntas de cada uno de los niveles de lectura descritos anteriormente, elemental, intermedio y avanzado, que se pueden responder a partir de la **Figura III.42**.

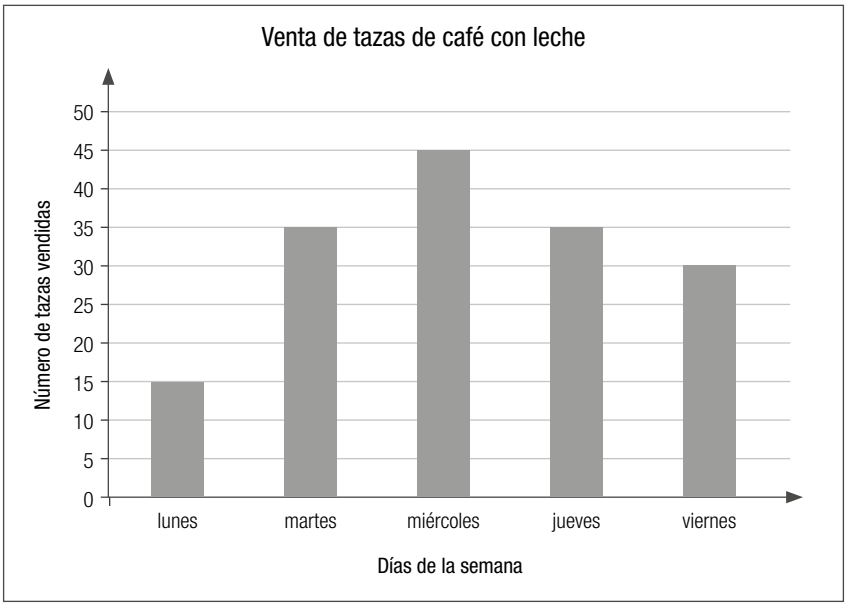
Nivel de lectura	Pregunta	Respuesta
Elemental	¿Para qué años el gráfico muestra información respecto del Índice de Precios al Consumidor?	Para 2010, 2011 y 2012.
Intermedio	Si consideramos cada año por separado, ¿en cuál de ellos creció más el Índice de Precios al Consumidor desde enero a diciembre?	En 2011.
Avanzado	Dado el comportamiento del Índice de Precios al Consumidor en 2010, 2011 y 2012, se cree que los valores que este alcance durante 2013 serán mayores a los de 2012. Justifique esta afirmación. De acuerdo al comportamiento del Índice de Precios al Consumidor en 2010, 2011 y 2012, ¿cómo cree usted que se comportará este índice durante los meses de junio a agosto de 2013?	En 2011 y 2012, el Índice de Precios al Consumidor ha sido mayor que en su año precedente, dado que su curva asociada se encuentra sobre la del año anterior. Si esta tendencia se mantiene, es esperable que este índice alcance mayores valores en 2013 que en 2012. Es posible que el Índice de Precios al Consumidor muestre un crecimiento más lento o nulo durante dichos meses de 2013, dado que es el patrón observado en 2010, 2011 y 2012.

Tabla III.44: Algunas preguntas que pueden responderse a partir del gráfico de la Figura III.41, y los niveles de lectura a los que corresponden.

Ejercicios

1. Los gráficos muestran la venta de sopaipillas y cafés con leche de un quiosco. Cada sopaipilla se vende a $100 y cada taza de café con leche se vende a $100.





a. ¿A qué nivel de lectura pertenecen las siguientes preguntas? Justifique su respuesta utilizando como argumento las tareas que deberán desarrollar los alumnos para responderlas.

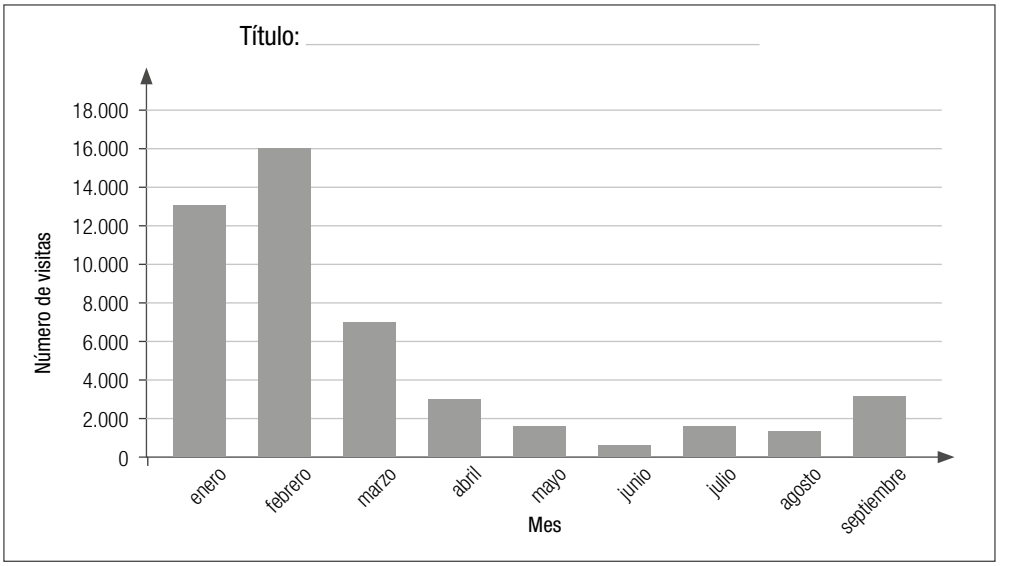
   i. ¿Es más conveniente para el quiosco vender sopaipillas o café con leche? Justifique su respuesta.

   ii. ¿Cuánto dinero obtiene el quiosco de la venta semanal de sopaipillas?

b. Complete la siguiente tabla con, al menos, un ejemplo de pregunta de cada nivel de lectura que pueda responderse a partir de la figura.

Nivel de lectura	Pregunta	Respuesta
Elemental		
Intermedio		
Avanzado		

2. La Cueva del Milodón es un monumento natural ubicado en la XII Región, en la zona austral de nuestro país. En este lugar fueron encontrados restos de dicho animal prehistórico. El gráfico muestra la cantidad de visitantes a ese parque, en algunos meses del año 2008.



a. Indique un título adecuado para el gráfico. ¿Qué nivel de lectura de gráficos es necesario usar para completar esta tarea?

b. ¿Cuál cree usted que será el comportamiento del número de visitantes a la Cueva del Milodón durante los meses de octubre a diciembre? Indique el nivel de lectura necesario para responder esta pregunta.

4. Elección del tipo de representación

Hasta aquí, hemos visto formas de resumir y presentar la información contenida en un conjunto de observaciones que pueden clasificarse en tablas de frecuencias y gráficos. Debemos, entonces, preguntarnos las consideraciones que se deben tener en cuenta para determinar qué representación usar para comunicar la información contenida en los datos.

Para responder la pregunta anterior, se debe tener en cuenta las siguientes ideas claves:

- Tanto tablas como gráficos tienen distintas potencialidades para resumir o comunicar información.
- Ambos, tablas y gráficos permiten obtener información cuantitativa sobre el conjunto de datos; sin embargo, usualmente es más fácil obtenerla a partir de una tabla.
- Un gráfico permite acceder no solo a información cuantitativa sobre el conjunto de datos, como leer la frecuencia de una categoría, sino también a información cualitativa, como la forma de un histograma, mediante la visualización e interpretación de los componentes que los caracterizan.

La **Tabla III.45** corresponde a un cuadro comparativo que puede ayudar a decidir qué representación utilizar. Sin embargo, recalcamos que ambos tipos de representaciones son en realidad complementarios.

	Gráficos	Tablas
¿Qué permite comunicar?	Información cuantitativa y cualitativa sobre los datos. Su principal fortaleza es esta última.	Información cuantitativa exacta sobre los datos.
¿Bajo qué condiciones se debe usar?	Para comunicar información relacionada a tendencias o patrones en los datos.	Para comunicar información que requiera de cantidades exactas. Es posible identificar patrones a partir de una tabla, sin embargo, suele ser más engorroso.
¿Qué acciones permite ejecutar?	Tomar decisiones en base a comparaciones o apreciaciones generales.	Tomar decisiones en base a operaciones con resúmenes numéricos de los datos.

Tabla III.45: Comparación entre características de representaciones gráficas y de tablas.

Entonces, el tipo de representación elegida depende de: la información que se quiera entregar al observador o receptor; las preguntas que se quieran responder, en el caso de datos que surgen en el marco de preguntas de investigación; y la madurez cognitiva de los individuos receptores y/o creadores de la representación, entre otros.

A manera de ejemplo, para comunicar información exacta, es preferible usar tablas de frecuencias, mientras que para estudiar una tendencia, es preferible utilizar un gráfico.

Según lo que hemos estudiado en este capítulo, entonces, la representación utilizada debe considerar la naturaleza de los datos, cualitativos o cuantitativos, y la información que se desee comunicar. En particular, recordemos que esta información debe relacionarse con el problema que se plantee, es decir, con la pregunta de investigación que dio origen al estudio.

Ejercicios del capítulo

1. Determine qué permite comunicar cada uno de los siguientes gráficos. Justifique su respuesta.
   a. Gráfico concreto.
   b. Pictograma.
   c. Gráfico de barras vertical.
   d. Gráfico de barras horizontal.
   e. Gráfico circular.

2. Para cada una de las siguientes preguntas de investigación, sugiera gráficos apropiados, de acuerdo al tipo de variable considerada.
   a. ¿Cuáles son los deportes favoritos del segundo básico?
   b. ¿Quiénes tienen las mejores calificaciones en Educación Física en la escuela?
   c. ¿Qué comida saludable consumen los alumnos de la escuela?
   d. ¿En qué lugar de la escuela prefieren pasear los alumnos?
   e. ¿Cuáles son los programas de televisión favoritos de las madres de los alumnos de primero básico?
   f. ¿Cuántas calorías queman los alumnos durante un día de jornada escolar? ¿cuál es la diferencia con las que queman en un día del fin de semana?

3. Los siguientes datos corresponden a las calificaciones finales de un curso:

60	58	40	51
53	62	48	68
47	62	38	47
45	44	62	47
47	37	58	32
50	40	35	65

a. Determine cuál(es) de los siguientes gráficos puede(n) ser utilizados para representar estas calificaciones. Justifique su respuesta.

Gráfico concreto.

Pictograma.

Gráfico de barras vertical.

Gráfico de barras horizontal.

Gráfico circular.

b. Construya todos los gráficos que sean pertinentes, de acuerdo a su respuesta en el apartado anterior.

c. Complete la siguiente tabla y responda los apartados que siguen.

Calificaciones	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa porcentual	Frecuencia relativa porcentual acumulada
1,0-1,9				
2,0-2,9				
3,0-3,9				
4,0-4,9				
5,0-5,9				
6,0-7,0				

d. ¿Cuál es la clase con la menor frecuencia?

e. ¿Qué porcentaje de alumnos obtuvo una nota bajo 5,0?

f. ¿Qué porcentaje de alumnos obtuvo nota 4,0 o superior?

4. Elisa realizó una encuesta para conocer las mascotas favoritas de sus compañeros de curso. Ella construyó una tabla con los datos que recolectó, pero se le rompió la hoja y perdió información. El trozo de hoja que Elisa pudo rescatar fue:

Mascota	Frecuencia	Frecuencia Relativa
Perro	8	$\frac{1}{4}$
Gato	6	
Pez	4	
Hámster	2	
Otras / No tiene		

¿Cuántos alumnos marcaron la preferencia Otras/No tiene? Justifique su respuesta.

5. Para cada una de las siguientes tablas, indique el tipo de representación gráfica que utilizaría para mostrar la información. Construyalas y refiérase a las principales características de los datos que evidencian estas representaciones.

Estatura de Luis, en el día de su cumpleaños.

Año	Estatura (m)
2006	1,32 m
2007	1,35 m
2008	1,42 m
2009	1,45 m

Votación para elegir al mejor compañero en un curso

Alumno	Votos
Roberto	7
Manuela	8
Feliciano	3
Marcela	10
Carlos	2

Indique, en cada caso, el tipo de gráfico que usaría para la representación de la información. Justifique su elección.

Capítulo
IV

Medidas o estadísticos de resumen

Introducción

Como vimos en el capítulo anterior, un conjunto de datos puede ser representado a través de tablas de frecuencias y gráficos. La principal característica de estas representaciones es que, además de entregar información exacta o cuantitativa, como frecuencias absolutas, o relativas simples o porcentuales, nos entregan información visual de tipo cualitativa. A modo de ejemplo, en un histograma podemos visualizar intervalos de valores donde se encuentran la mayor parte de las observaciones o, en particular, notar que las observaciones están concentradas en valores altos de la variable, habiendo pocas observaciones de valores pequeños, o viceversa. Podemos, también, comparar los histogramas de dos conjuntos de datos diferentes y notar, por ejemplo, que sus formas son similares, pero que se mueven en intervalos diferentes de valores, entre otros aspectos.

Notamos que en afirmaciones como las anteriores nos referimos a características específicas sobre la distribución de los datos, como intervalos de valores más frecuentes o localización de la distribución. Algunas de estas características pueden ser cuantificadas a través de las llamadas *medidas* o *estadísticos de resumen*. En particular, estudiaremos medidas, o estadísticos, de tendencia central, de posición relativa y de dispersión.

Este capítulo está organizado como sigue: la **Sección 1** muestra la importancia de entender e interpretar correctamente las medidas de resumen de distribuciones de datos. Las **Secciones 2**, **3** y **4** se enfocan, cada una, en estadísticos que representan aspectos diferentes de un conjunto de observaciones. En particular, la **Sección 2** trata sobre medidas de tendencia central y se estudian los conceptos de media, mediana y moda. La **Sección 3** se concentra en medidas de posición relativa estudiando, en particular, cuartiles, quintiles, deciles y percentiles, y finalizando con el diagrama de cajón con bigotes, o *boxplot*, que integra estos estadísticos. Finalmente, la **Sección 4** estudia medidas de dispersión o variabilidad de los datos.

1. Motivación

Muy frecuentemente encontramos estudios o reportes que describen conjuntos de datos de manera muy sucinta, y entregan, entre otras características, algunos promedios, medianas o desviaciones estándar. Si bien el ciudadano medio posee alguna idea de lo que estos valores representan, resulta importante comprender exactamente su significado, de modo de extraer conclusiones adecuadas, que tomen en cuenta las limitaciones que estas medidas puedan tener. Por otra parte, muchos profesionales deben comunicar hallazgos o resultados de sus estudios o mediciones, para lo cual es necesario que entiendan qué medidas son relevantes en cada caso, sepan la manera de obtenerlas y sean capaces de elaborar conclusiones en torno a los valores obtenidos.

Consideremos, a modo de ejemplo, los resultados del SIMCE, que suelen reportarse a través de una tabla como la **Tabla IV.1**, que muestra los resultados a nivel nacional de las pruebas de Lenguaje, Matemática y Ciencias Naturales de los cuartos básicos, en el año 2011.

Prueba	Puntaje promedio 2011
Lenguaje	267
Matemática	259
Ciencias Naturales	259

Tabla IV.1: Resultados de las pruebas SIMCE de Lenguaje, Matemática y Ciencias Naturales a nivel nacional, de los cuartos básicos, en el año 2011.

En la tabla, se leen los promedios de los puntajes de todos los niños del país en cada una de las evaluaciones. Leemos, por ejemplo, que el promedio nacional en la prueba de Lenguaje fue de **267** puntos.

Sin embargo, únicamente a partir de la tabla no podemos responder preguntas como:

- ¿Se puede afirmar que hay igual cantidad de puntajes sobre y bajo el promedio de los resultados nacionales?
- ¿Qué tan cercanos son, en general, los puntajes de cada prueba a su promedio nacional?
- ¿Son los promedios nacionales de los puntajes, buenos representantes de los puntajes de todos los niños del país que rindieron la prueba? ¿en qué sentido?
- Los promedios nacionales en las pruebas de Matemática y Ciencias Naturales son los mismos. ¿Significa esto que los niños dominan sus contenidos en igual medida?

En este capítulo, entregaremos herramientas que nos ayudarán a responder preguntas como las anteriores.

2. Medidas de tendencia central

Consideremos un grupo de niños que juega cada uno con un número diferente de cubos, como se muestra en la **Figura IV.1**. Si registramos el número de cubos con que juega cada niño obtenemos:

4 1 5 5 0

Figura IV.1: Cada niño juega con un número diferente de cubos.

¿Existe algún número de cubos que sea representativo del número de cubos con que juegan los niños? Aunque en este ejemplo se cuenta con muy pocos datos, sabemos que en el caso de variables cuantitativas podemos representar estos números tanto en una tabla de frecuencias, como en un histograma. A partir de dichas representaciones, sería posible determinar, por ejemplo, que el número de cubos más frecuente es 5. Sin embargo, existen también otros conceptos de *valor representativo*. En efecto, podemos visualizar, al menos, 3 posibles respuestas:

- 5 cubos es el número de cubos más frecuente.
- El promedio de cubos de los niños es 3.
- Si ordenamos las observaciones de menor a mayor, el número de cubos que queda en el centro es 4.

Las tres afirmaciones entregan una idea de valores representativos del número de cubos con que juegan los 5 niños. ¿Cuál de ellas debemos considerar? Las características que se han utilizado para responder a la pregunta se denominan *medidas de tendencia central*. A continuación estudiaremos cada una de ellas: la media o promedio, la mediana y la moda, y notaremos que cada una de ellas tiene una interpretación diferente, por lo que es importante saber lo que estamos comunicando al reportar sus valores.

2.1. La media o promedio

En el ejemplo de la **Figura IV.1**, una manera de encontrar un valor de resumen es a través de lo que se denomina *reparto equitativo*. Esto es: si juntamos todos los cubos de los niños y los repartimos entre ellos, asegurándonos de que cada niño reciba la misma cantidad de cubos, la media o promedio corresponde a la cantidad de cubos que recibiría cada niño.

En la **Figura IV.1** vemos que, en total, los niños tienen **15** cubos, pero que no todos juegan con el mismo número de estos. En la **Figura IV.2**, se han repartido de manera equitativa estos **15** cubos, de modo que todos los niños juegan ahora con **3** cubos. Así, decimos que la media o promedio de los cubos de los niños es **3** cubos.

Figura IV.2: Los cubos han sido repartidos de manera equitativa, y ahora cada uno de los niños tiene 3 cubos. Luego, la media o promedio de la cantidad de cubos es 3.

Esta idea sugiere una forma de calcular la media: para repartir los cubos, primero podemos juntarlos para saber cuántos cubos debemos repartir. Aritméticamente, esto significa que debemos sumar los números de cubos que tienen los niños. En este caso, obtenemos un total de $4 + 1 + 5 + 5 + 0 = 15$ cubos. Una vez que sabemos cuántos cubos se tienen en total, debemos repartirlos de tal manera que cada uno de los 5 niños reciba la misma cantidad de cubos. Si bien existen diferentes estrategias para repartir los cubos, como, por ejemplo, repartirlos uno a uno siguiendo algún orden en los niños de manera consecutiva, hasta que se nos acaben los cubos, para entender el concepto de media utilizaremos la estrategia de calcular primero cuántos cubos corresponden a cada niño y luego repartir esta cantidad a cada uno, de una sola vez. Esto significa que debemos dividir la suma obtenida por el total de niños, es decir, realizar la operación $\frac{15}{5} = 3$ cubos.

En resumen, el cálculo realizado corresponde a:

$$\frac{4+1+5+5+0}{5}=3$$

En el caso general en que disponemos de un conjunto de n observaciones, que anotamos como $x_1, x_2, \ldots, x_n$, podemos obtener la media de la distribución de las observaciones, que se representa por el símbolo $\bar{x}$, como:

$$\bar{x}=\frac{x_1+x_2+\ldots+x_n}{n}$$

En el ejemplo que planteamos sobre la cantidad de cubos que tiene cada niño de la **Figura IV.1**, partiendo con el niño de arriba a la izquierda, $x_1=4, x_2=1, x_3=5, x_4=5, x_5=0$ y $n=5$, y la fórmula recién descrita corresponde a

$$\bar{x}=\frac{4+1+5+5+0}{5}=3$$

Esta operación ya la habíamos realizado con anterioridad. Notemos que dado que la suma es conmutativa, el orden en que recolectamos las observaciones no es relevante.

Dijimos anteriormente que, al pensar en repartir los cubos a los niños, no estamos pensando en la repartición de los cubos uno a uno. En efecto, supongamos que, en total, los niños tuviesen 12 cubos. Si los repartimos uno a uno siguiendo un orden preestablecido de los niños, al entregar el décimo cubo habremos entregado un total de 2 cubos a cada uno de los 5 niños. Al repartir los 2 cubos restantes a los 2 niños siguientes según el ordenamiento, tendremos 2 niños con 3 cubos y 3 niños que tendrán solo 2 cubos. No habremos logrado el objetivo del reparto equitativo.

El ejemplo recién mencionado muestra también que, si bien la interpretación de la media a través del reparto equitativo es una de las más intuitivas, presenta ciertas limitaciones. En efecto, en el caso anterior, la media del número de cubos corresponde a $\frac{12}{5}=2{,}4$ cubos. ¿Cómo pueden repartirse 2,4 cubos a cada niño, si no podemos particionar los cubos? Esta es una de las características de la media: ella puede tomar valores que no corresponden a valores válidos de la variable, lo cual genera problemas en una representación a través del concepto de reparto equitativo, cuando los objetos no son particionables, como es el caso de los cubos. En situaciones como esta, no es aconsejable utilizar la interpretación de reparto equitativo para la media.

Otra manera de interpretar la media es a través del concepto de nivelación de los valores de las observaciones. A modo de ejemplo, consideremos 3 varillas de alturas 30, 20 y 40 centímetros, y supongamos que las ubicamos una al lado de la otra, como se muestra en la **Figura IV.3**, arriba. Como las 3 varillas tienen diferentes alturas, no sería posible apoyar un objeto sobre ellas de manera perfectamente horizontal. Si cortamos ahora la varilla de 40 centímetros en dos trozos de 30 y 10 centímetros, respectivamente, como muestra la **Figura IV.3**, al centro, y añadimos el trozo de 10 centímetros a la varilla de 20, como se muestra en la **Figura IV.3**, abajo, las 3 varillas resultantes tendrán una altura de 30 centímetros, que corresponde al valor de la media, y se podría apoyar un objeto sobre ellas de manera perfectamente horizontal. Esta idea puede visualizarse, por ejemplo, al considerar la construcción de una mesa. Si las patas a utilizar tienen diferentes alturas, la cubierta de la mesa quedará inclinada. Si pudiéramos realizar cortes y uniones a las patas, el procedimiento de nivelación descrito lograría que la cubierta de la mesa quede perfectamente horizontal y la altura de cada pata correspondería al promedio de sus alturas originales.

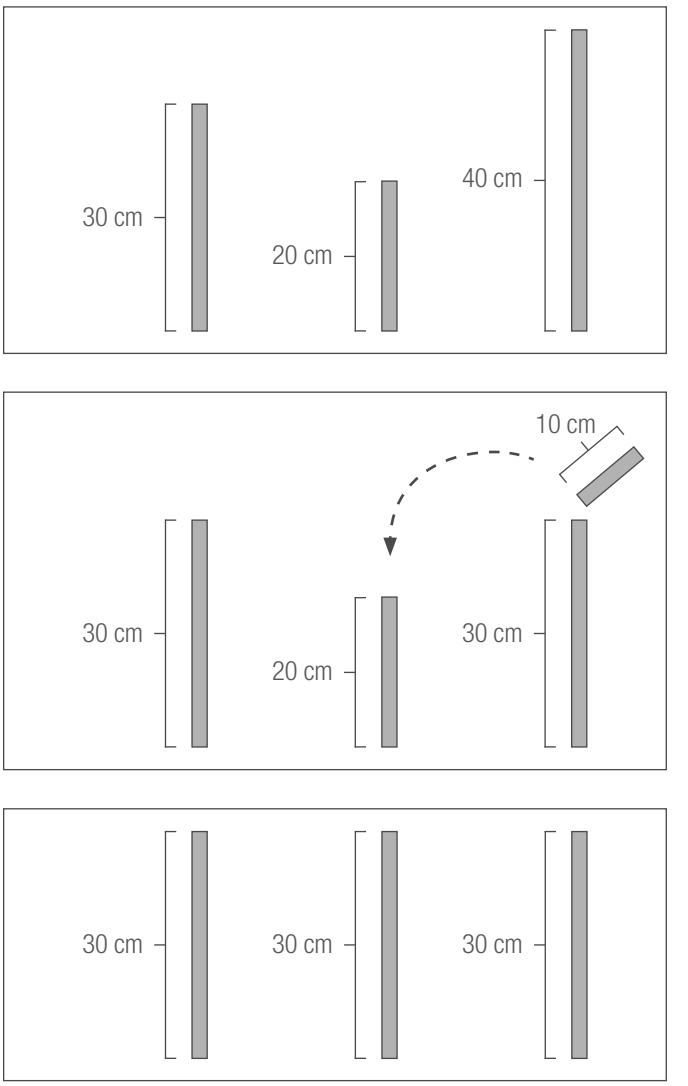


Figura IV.3: La media como nivelación. Arriba: las varillas tienen diferentes alturas. Centro: quitamos un trozo de 10 cm a la varilla más alta. Abajo: unimos este trozo a la varilla más baja. Las tres varillas resultantes tienen ahora la misma altura, 30 cm. Esta altura común corresponde a la media de las alturas de las barras.

Otra manera de interpretar la media corresponde a un concepto de igualdad de sumas de distancias, idea que ilustraremos a través de un ejemplo. Consideremos las notas de 4 niños de un curso:

3,0 7,0 7,0 7,0

Al realizar las operaciones, encontramos que la media de este conjunto de datos corresponde a un **6,0**. Veremos que esta nota es la única nota tal que la suma de las distancias desde sí misma a las observaciones que son menores a ella es igual a la suma de las distancias desde sí misma a las observaciones que son mayores.

Para ayudar a la comprensión, comenzaremos considerando solo 2 niños: el niño que obtuvo la nota **3,0**, y uno de los 3 niños que obtuvo nota **7,0**. La media de las notas de estos 2 niños es **5,0**. La **Figura IV.4** muestra las 2 observaciones consideradas, a través de círculos en los valores **3** y **7**, y sus distancias a la nota **5,0**. La distancia desde la observación menor que **5,0** y este último es $|3-5|=2$, y la distancia desde la observación que es mayor que **5,0** y este último es $|7-5|=2$, logrando que las distancias a ambos lados de la media, **5,0**, sean las mismas, como habíamos adelantado.

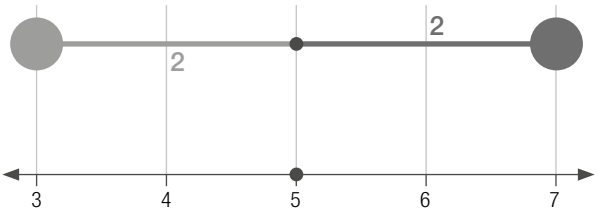


Figura IV.4: La distancia entre la notas 3,0 y 5,0 es $|3-5|=2$, la misma que la distancia entre las notas 7,0 y 5,0, $|7-5|=2$.

Incorporemos ahora a los 2 niños restantes que obtuvieron un **7,0**. La **Figura IV.5** muestra las 4 observaciones mediante círculos, y sus distancias a la nota **5,0**. La distancia desde la nota menor que **5,0** y esta última es:

$$|3-5|=2$$

Mientras que las distancias desde las notas mayores que **5,0** y esta última es:

$$|7-5|=2 \qquad |7-5|=2 \qquad |7-5|=2$$

Cuya suma es **6**. Luego, la suma de las distancias de las observaciones menores que **5,0**, que es **2**, no es igual a la suma de las distancias de las observaciones que son mayores, que es **6**. Luego, la nota **5,0** ya no cumple con la condición pedida.

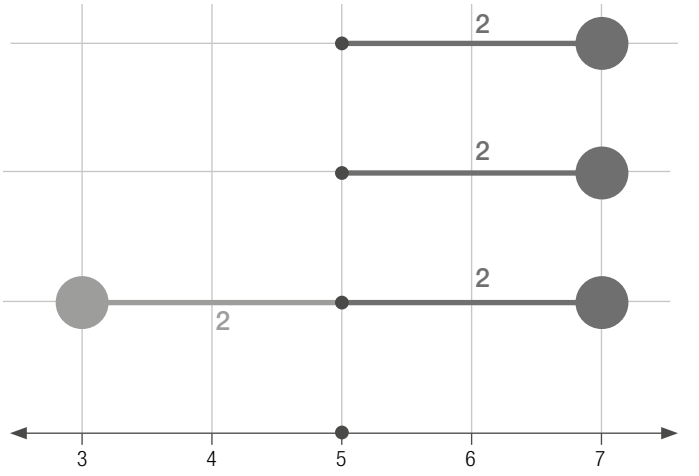


Figura IV.5: La distancia desde la nota menor a 5,0 es $|3-5|=2$. Las distancias desde las tres notas mayores que 5,0 son todas iguales a $|7-5|=2$, por lo que su suma, 6, es diferente a la anterior.

Dado que la suma de las distancias de las observaciones mayores que la nota **5,0** es mayor, podemos intuir que debemos mover el punto buscado hacia la derecha, con el objeto de disminuir las distancias desde este a las observaciones mayores. Veremos que esto realmente ocurre. En efecto, notemos que, algebraicamente, lo que se quiere es encontrar un punto a tal que:

$$|3 - a| = |7 - a| + |7 - a| + |7 - a|$$

Asumiendo que el punto buscado es mayor que **3,0** y menor que **7,0**, podemos eliminar los valores absolutos en la expresión de la manera:

$$a - 3 = (7 - a) + (7 - a) + (7 - a)$$

Es decir, $-3 - 7 - 7 - 7 + 4a = 0$. Despejando el valor de a, obtenemos que este debe ser igual a:

$$a = \frac{3 + 7 + 7 + 7}{4} = 6$$

lo que coincide con la media de las notas de los 4 niños.

Esta situación se ilustra en la **Figura IV.6**, que muestra que la distancia desde la observación menor que la media, que es **3**, es igual a la suma de las distancias desde las observaciones mayores que la media, que también es $(1 + 1 + 1) = 3$.

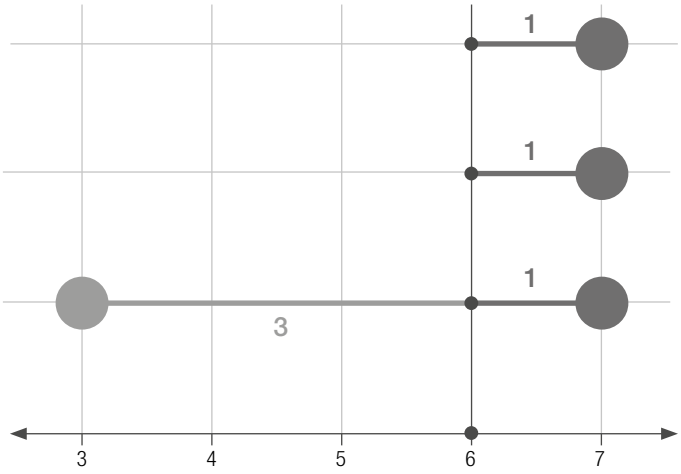


Figura IV.6: La distancia entre la nota menor que la media y esta última, que es 3, es igual a la suma de las distancias entre las 3 notas mayores que la media y esta última, que es (1+1+1 = 3).

Notemos, que en el ejemplo que seguimos, solo existen 2 valores en el conjunto de datos, **3,0** y **7,0**. Se puede mostrar que la condición de igualdad de sumas de distancias se cumple en cualquier conjunto de observaciones, aunque existan más de dos valores diferentes en él. El siguiente ejemplo ilustra un caso como este.

Ejemplo

Suponga que las notas de otros 4 alumnos son:

$$4{,}0 \qquad 5{,}0 \qquad 5{,}5 \qquad 6{,}3$$

Cuya media es 5,2. La suma de las distancias desde las observaciones menores a 5,2 a este punto es:

$$|\,4 - 5{,}2\,| + |\,5{,}0 - 5{,}2\,| = 1{,}4$$

Y la suma de las distancias desde las observaciones mayores a 5,2 a este punto es:

$$|\,5{,}5 - 5{,}2\,| + |\,6{,}3 - 5{,}2\,| = 1{,}4$$

Lo que es igual a la distancia anterior.

Se puede demostrar que, en todo conjunto de observaciones, la media corresponde al único valor que cumple con la propiedad pedida.

De este modo, en general, podemos interpretar la media como el punto tal que la suma de las distancias desde las observaciones a su izquierda y este mismo es igual a la suma de las distancias desde las observaciones a su derecha y este mismo.

Notemos que, a diferencia de lo que ocurre con la interpretación de reparto equitativo cuando los elementos a repartir no son particionables, tanto la interpretación de la media como nivelación, como la de igualdad de sumas de distancias pueden utilizarse sin necesidad de que la media tome valores enteros. En efecto, las barras verticales pueden nivelarse a cualquier altura, así como también el punto que cumple la condición de igualdad de sumas de distancias puede encontrarse en cualquier ubicación entre los datos.

En resumen

- La *media* o *promedio* corresponde a una medida de tendencia central, que puede ser interpretada en base a los conceptos de reparto equitativo, nivelación e igualdad de sumas de distancias.
- El concepto de *reparto equitativo* sugiere la manera de obtener el valor de la media: "juntamos" (sumamos) los valores de las observaciones como si fueran unidades, y las repartimos equitativamente (dividimos) entre todas ellas.
- La media de un conjunto de datos puede no ser un valor admisible de la variable de interés, situación que dificulta su interpretación a través del concepto de reparto equitativo.

Notemos que, muchas veces, se presenta la media a través de las operaciones aritméticas necesarias para obtener su valor. A modo de ejemplo, suele encontrarse que "la media corresponde a la suma de los valores de las observaciones, dividida por el número total de ellas". Sin embargo, el énfasis debe estar en comprender lo que el concepto de media representa. Hemos mostrado aquí

3 interpretaciones importantes: reparto equitativo, nivelación e igualdad de sumas de distancias, que explican de qué forma la media es un valor representativo de los valores de las observaciones.

Notemos que solo tiene sentido hablar de la media, o promedio, cuando los datos son cuantitativos. Al igual como se hizo al describir variables cualitativas en el **Capítulo 3**, hacemos hincapié en que no se deben confundir las variables cualitativas que han sido codificadas con etiquetas numéricas como, por ejemplo, rojo = **1**, azul = **2**, etc., con variables cuantitativas. Esto también es válido para las variables cualitativas ordinales, como grados de acuerdo y desacuerdo. En efecto, aunque por tener categorías ordenables podríamos asignar números a sus valores, no existe el concepto de distancia entre categorías que determine las distancias entre los valores asignados.

2.2. La mediana

Supongamos que nos interesa estudiar la altura de los niños del quinto básico de una escuela. Podemos encontrar un concepto de centro de las alturas ubicando a los niños en una fila en orden creciente (o decreciente) de altura y encontrando al niño que queda en la posición central. Notamos que la altura de dicho niño corresponde a una característica del centro de las alturas de los niños. De este modo, llamamos *mediana* de la altura de los niños del curso, a la altura de la niña ubicado en la posición central, como se ilustra en la **Figura IV.7**.

Figura IV.7: Los niños han sido ordenados desde el más pequeño al más grande. La mediana corresponde a la altura de la niña central.

En otro ejemplo, consideremos las notas finales obtenidas por un grupo de **13** alumnos en un curso, las que han sido ordenadas de manera creciente:

4,3 4,6 4,6 4,7 5,0 5,0 **5,0** 5,0
5,7 5,9 6,1 6,1 6,6

La observación central corresponde al puntaje en negrita, puesto que existen **6** observaciones a su izquierda y **6** observaciones a su derecha. Luego, la mediana de las notas de este curso corresponde a **5,0**. Notemos que se obtiene el mismo resultado al ordenar las observaciones en orden creciente o decreciente, lo que ocurre con la mediana en todo conjunto de datos.

Tanto en el problema de las alturas de los niños, en la **Figura IV.7**, como en el de las notas finales de un curso, el número de observaciones en el conjunto de datos ha sido impar, 5 y 13, respectivamente. En este caso, al ordenar los valores de menor a mayor, es posible determinar la observación que está al centro dejando igual número de observaciones a su izquierda y a su derecha. Sin embargo, esto no es posible si el número de observaciones es par.

En efecto, consideremos nuevamente las edades, medidas en años, de los 34 profesores de una escuela, que tratamos en el **Capítulo 3**. Las edades han sido ordenadas de manera creciente y quedaron como sigue:

31	32	32	32	33	35	36	37	37	38
39	39	40	40	41	42	**42**	**43**	43	44
45	46	47	48	48	51	53	55	56	56
			60	61	62	76			

En este caso, no existe una única observación que esté al centro. En efecto, ambas observaciones en negrita, 42 y 43 años, corresponden a observaciones centrales, en el sentido de que hay 16 observaciones a la izquierda de 42, y 16 observaciones a la derecha de 43. Debemos extender el concepto de mediana, para lo que diremos que corresponde a un valor, no necesariamente una observación en el conjunto de datos, que deja el mismo número de observaciones a su izquierda que a su derecha. El valor 42 años no cumple con la condición, puesto que deja a su izquierda 16 observaciones y, a su derecha 17, dos cantidades diferentes. Algo similar ocurre con la observación 43 años.

Según el concepto anterior, la mediana deja de ser única. Efectivamente, en el ejemplo, cualquiera de los infinitos valores entre 42 y 43 años cumple con esta propiedad. Si tomamos como ejemplo el valor 42,7 años, este deja el mismo número de observaciones a su izquierda, que es 17, y a su derecha. En estos casos, lo habitual es convenir que la mediana corresponda al promedio entre las observaciones centrales. En el ejemplo se obtiene:

$$\text{Mediana} = \frac{42 + 43}{2} = 42{,}5 \text{ años}$$

Notamos que, al igual que con la media, la mediana puede no corresponder a un valor posible de la variable de interés.

El diagrama en la **Figura IV.8** resume el procedimiento para encontrar la mediana, tanto cuando el número de observaciones es par, como impar.

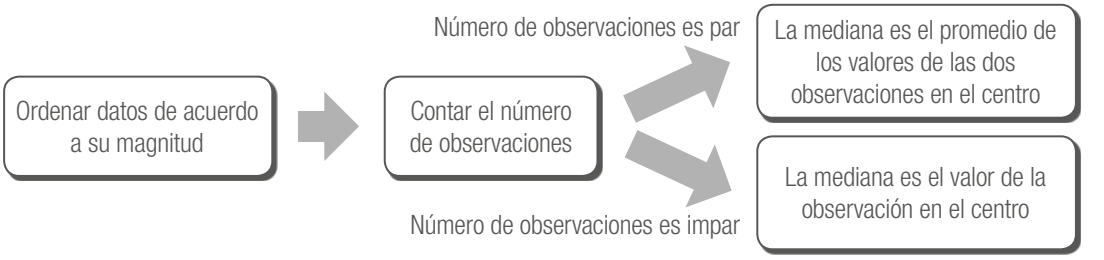
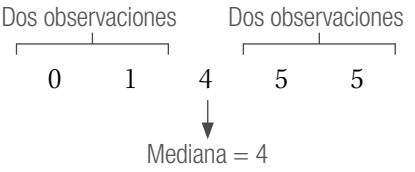


Figura IV.8: Esquema del procedimiento para obtener el valor de la mediana de un conjunto de observaciones.

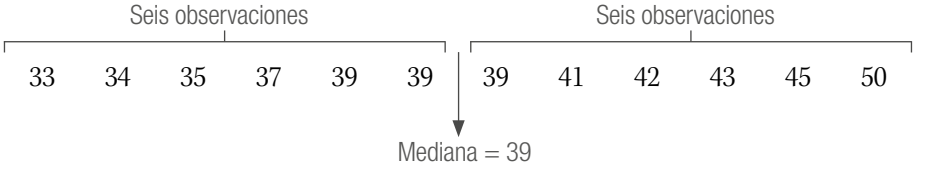
Hasta aquí hemos introducido la mediana a través de la idea de *valor central*. Al igual que con la media, veremos también otras interpretaciones de la mediana. En particular: de la idea de valor central, se desprende que *la mediana es un punto que divide a las observaciones en dos grupos de igual tamaño*, y por otra parte, *la mediana es tal que, al menos, el 50% de las observaciones es menor o igual a ella y, al menos, el 50% de las observaciones es mayor o igual que la misma.*

Derivaremos ambas interpretaciones a través de un ejemplo. Para la primera de ellas, consideremos nuevamente el problema de estudiar la cantidad de cubos que tiene cada uno de los 5 niños en la **Figura IV.1**. Si ordenamos los números de cubos de menor a mayor, obtenemos:



Donde la mediana corresponde al dato central, 4 cubos. Vemos que la mediana divide a las observaciones en 2 grupos, cada uno con 2 observaciones.

Consideremos también los siguientes datos, que corresponden a las edades de las madres de 12 niños en un Taller de Guitarra:



Según la convención que tomamos, la mediana en este caso corresponde al promedio de las dos observaciones centrales. Por ser ambas iguales a 39 años, su promedio corresponde también a 39 años. Vemos que la mediana dividió a las observaciones en dos grupos, a su izquierda y a su derecha, y que nuevamente, estos contienen el mismo número de observaciones que, en este caso, es 6.

Esto nos da entonces la interpretación de la mediana como un punto que divide al conjunto de observaciones, al ordenarlas en orden creciente (o decreciente) de magnitud, en dos grupos de igual tamaño. Notemos que esto ocurre tanto cuando el número de observaciones es par como impar, según vimos en los dos ejemplos recientes.

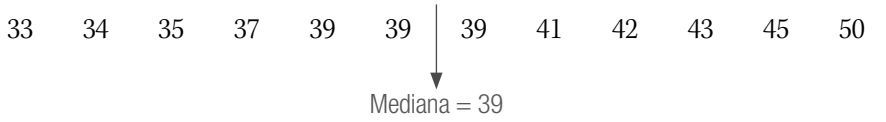
Para la segunda interpretación, *la mediana es tal que, al menos, el 50% de las observaciones es menor o igual a ella y, al menos, el 50% de las observaciones es mayor o igual ella*, consideremos nuevamente el número de cubos de los 5 niños:

0 1 4 5 5

Donde la mediana es 4 cubos. En este caso, las observaciones menores o iguales a la mediana son 3: 0 cubos, 1 cubo y 4 cubos. Es decir, $\frac{3}{5}$ de las observaciones, o un 60%, son menores o iguales a la mediana. Por otra parte, las observaciones mayores o iguales a la mediana también son 3: 4 cubos, 5 cubos y 5 cubos. Es decir, un 60% de las observaciones es mayor o igual a la mediana.

Vemos que en este conjunto de datos la mediana cumple con la propiedad deseada, dado que ambos porcentajes son mayores a 50%.

En el ejemplo anterior, ocurre que los porcentajes referidos son iguales, lo cual no se cumple en todo conjunto de observaciones. En efecto, consideremos nuevamente las edades de las madres:



Vimos que la mediana corresponde a 39 años. En este caso, el número de observaciones menores o iguales a la mediana es 7, lo cual corresponde a $\frac{7}{12}$ de las observaciones o, aproximadamente, 58%. Por otra parte, el número de observaciones mayores o iguales a la mediana es 8, lo cual corresponde a $\frac{8}{12}$ de las observaciones o, aproximadamente, 67%. Vemos que, en este caso, la mediana cumple con las condiciones pedidas, dado que ambos porcentajes son mayores a 50%; sin embargo, los porcentajes involucrados no son los mismos. Esto se debe a que el conjunto de datos contiene más de una observación que toma el valor de la mediana, y a que los 2 grupos formados contienen diferente número de estas repeticiones: 2 y 1 valores iguales a 39 en cada grupo. Este es el caso general.

Consideremos ahora un tercer ejemplo, que se refiere a los puntajes SIMCE de Lenguaje de cierto nivel y año, de todos los alumnos del país.

Para pensar

Si se nos indica que "la mediana de todos los puntajes de los niños es 260 puntos". ¿Tiene importancia, en este caso, que distingamos si los porcentajes de observaciones menores o iguales a 260 puntos y de observaciones mayores o iguales a 260 puntos son exactamente iguales? Justifique su respuesta.

En este ejemplo, el número de alumnos que rinde la prueba SIMCE cada año es considerablemente grande y es esperable que el número de observaciones iguales a la mediana sea relativamente pequeño, en comparación al número total de estos alumnos. Distinguir, en este caso, diferencias en los porcentajes carece de sentido, y podemos decir que "aproximadamente, un 50% de las observaciones es menor o igual que la mediana", y que "aproximadamente, un 50% de las observaciones es mayor o igual que la mediana".

En resumen

- La *mediana* corresponde a un valor que queda al centro al ordenar las observaciones creciente o decrecientemente. Si el número de observaciones es par, no existe un único valor "al centro" y, en ese caso, se propone elegir la mediana como el promedio de los valores de las dos observaciones centrales.
- La mediana también puede interpretarse como un valor que divide al conjunto de observaciones, ordenado según su magnitud, en 2 grupos de igual tamaño.
- La mediana también puede interpretarse como un valor tal que, al menos un **50%** de las observaciones son menores o iguales a ella y, al menos, un **50%** de las observaciones son mayores o iguales a ella.

Notamos que, en general, la mediana solo tiene sentido para observaciones o variables cuantitativas. Si bien es posible ordenar los valores de variables cualitativas ordinales, en general, estas variables presentan muy pocas categorías, por lo que la mediana carece de utilidad y no se recomienda su uso.

Ejercicios

1. Usted desea explicar a sus alumnos que la tasa de natalidad del país en el año 2011 fue de **1,87** infantes nacidos por mujer. ¿Qué interpretación o representación de la media utilizaría para este fin?

2. Un carpintero dispone de 4 palos de madera que utilizará para construir una mesa. Estos palos miden **100**, **82**, **106** y **70 cm**, respectivamente. Suponiendo que él puede trozar la madera y añadir cortes, represente los palos como en la **Figura IV.3** y muestre una secuencia para lograr que todos los palos tengan la misma altura, de modo que la cubierta de la mesa quede perfectamente horizontal.

3. Los siguientes datos corresponden a las edades, en años, de las madres de los alumnos del sexto básico:

   35 34 37 33 38 45 40 41 50 43 39 42

   Obtenga la media y la mediana de estos datos, y comente la similitud o diferencia entre estos valores.

4. Encuentre la media y la mediana de las alturas (en centímetros) de los 7 alumnos en el Taller de Teatro de un curso:

   153,8 154,7 156,9 154,3 152,3 156,1 152,3

5. Se ha determinado que un exceso de la concentración de plomo en el aire puede producir efectos nocivos para la salud. En efecto, se ha determinado el valor de **1,5** microgramos por metro cúbico como nivel máximo tolerable. Los siguientes datos corresponden a concentraciones de plomo, en microgramos por metro cúbico, medidos en el aire, al interior de un edificio de una escuela:

5,4 1,1 0,42 0,73 0,48 1,1

   a. Obtenga la concentración de plomo media, o promedio, en estas **6** mediciones.
   b. Obtenga la mediana de la concentración de plomo en estas **6** mediciones.
   c. Sabemos que ambos valores, la media y la mediana, corresponden a valores representativos del conjunto de datos. ¿A qué cree usted que se debe la diferencia obtenida entre la media y mediana de este conjunto?

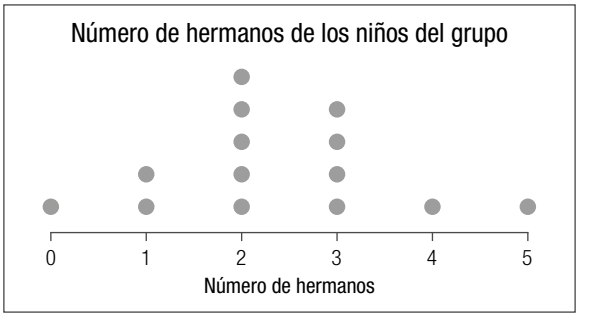
6. Para ejercitar el concepto de media, una profesora pide a los niños obtener el número promedio de lápices de colores que tiene cada uno de ellos en su estuche. Las siguientes son las cantidades que dice cada uno de los 15 niños del curso:

4 2 1 5 10 3 7 9
2 5 5 2 7 1 12

   a. Obtenga la media de este conjunto de datos.
   b. ¿De qué manera puede usted interpretar la media en este contexto, para explicarla a sus alumnos?
   c. Uno de los 15 niños se ha equivocado al decir que tiene **10** lápices, pues en realidad tiene **20**. Obtenga la media reemplazando la observación equivocada por la correcta, de **20** lápices. ¿Cómo cambia la media al compararla con su respuesta en a.?
   d. Obtenga la mediana de los datos originales reportados y luego, la mediana al corregir la observación de valor **10** lápices por **20** lápices. ¿Cómo cambia la mediana?

7. Uno de sus alumnos está interesado en conocer los resultados en la competencia del salto largo de su escuela. Un compañero le cuenta que el atleta que obtuvo el quinto lugar, entre **9** competidores, saltó una distancia de **2,15 m**. ¿A qué medida de tendencia central corresponde esta longitud?

8. Considere el siguiente diagrama de puntos sobre el número de hermanos de un grupo de niños:

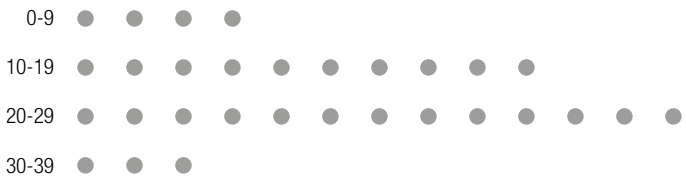


   a. Entregue una estimación de la media de este conjunto de observaciones, sin obtenerla de manera exacta, basándose en la interpretación de igualdad de sumas de distancias.

b. Obtenga la media de manera exacta y compare con su valor estimado en el apartado anterior.

c. Obtenga la mediana de este conjunto de observaciones.

9. Considere el siguiente diagrama de puntos que muestra las edades, en años, de los asistentes a una obra de teatro para todo público.



a. Entregue estimaciones de media y mediana de este conjunto de datos basándose únicamente en la figura (note que no puede obtener estos estadísticos de manera exacta).

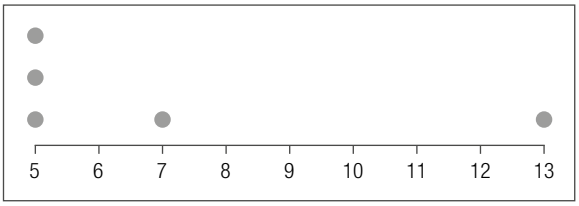
b. Suponga que se le indica que los valores exactos de las observaciones corresponden a:

4	5	8	9	10	10	12	13	14	14	14	16	18
18	20	20	21	21	21	24	25	26	26	27	28	28
29	31	34	35									

Obtenga ahora la media y la mediana del conjunto de datos. ¿Qué tan cerca o tan lejos están estos valores de los valores que usted estimó?

10. Suponga que un niño tiene en su estuche 4 lápices de distintos largos: **8 cm**, **8 cm**, **12 cm** y **16 cm**. Represente estos lápices como las barras en la **Figura IV.3** y obtenga el valor de la media a través del concepto de nivelación.

11. En el siguiente gráfico de puntos, determine el valor de la media utilizando la propiedad de igualdad de sumas de distancias. Dibuje las líneas de las distancias para apoyarse.



12. Supongamos que se le pregunta a un grupo de **12** niños por el número de horas que dedicaron la semana anterior a practicar un deporte. Las respuestas de los niños fueron:

1	2	2	3	3	3	3
	4	6	6	7	8	

a. Represente los datos en un gráfico de puntos.

b. Determine la media de los valores usando la noción de igualdad de sumas de distancias.

2.3. Comportamiento de la media y la mediana frente a observaciones extremas o atípicas

A modo de ejemplo, supongamos que las siguientes cantidades corresponden a los sueldos mensuales de 5 jugadores de básquetbol, en pesos, y que se desea entregar un valor que represente a estos sueldos.

100.000 100.000 100.000 100.000 4.000.000

Para pensar

¿Qué valor consideraría un mejor representante de las observaciones en este caso: la media o la mediana?

Vemos que 4 jugadores reciben sueldos iguales, de **$100.000** cada uno, y que solo un jugador recibe un sueldo mayor, bastante diferente a los anteriores, de **$4.000.000**. Podemos obtener la media de los sueldos que, expresada en pesos, corresponde a:

$$\frac{100.000 + 100.000 + 100.000 + 100.000 + 4.000.000}{5} = 880.000$$

Por otra parte, es posible obtener la mediana de los sueldos, que corresponde al valor central, es decir, **$100.000**.

Si bien ninguna medida por sí sola es capaz de representar la distribución de los sueldos en su totalidad, en este caso, parece mejor elegir el valor **$100.000** como representante, es decir, la mediana. En efecto, supongamos que un deportista debe decidir si dedicarse o no al básquetbol. En ese caso, el valor **$880.000** puede darle la idea equivocada de que podría obtener aproximadamente dicho sueldo, cuando, en realidad, hay altas chances de que reciba solo **$100.000**.

La situación observada ocurrió por la presencia de un valor extremo o atípico, correspondiente al sueldo de **$4.000.000**, mucho mayor que los sueldos restantes. En general, un valor extremo corresponde al valor de una observación que se aleja del grueso de los datos, ya sea porque es muy grande o muy pequeño. Existen algunas técnicas formales para determinar a qué magnitudes llamamos grandes o pequeñas con respecto a las observaciones restantes del conjunto, pero, por ahora, utilizaremos nociones intuitivas.

Para estos efectos, resulta útil considerar la representación de igualdad de sumas de distancias de la media. Consideremos los siguientes datos que representan el número de veces que un grupo de 5 personas asistió a una función de cine durante los últimos 6 meses:

6 3 6 2 3

Su promedio es 4 veces. La **Figura IV.9** muestra la representación que hemos utilizado de la media a través del concepto de igualdad de sumas de distancias. En efecto, la suma de las distancias desde las observaciones menores que 4 a este valor es $2 + 1 + 1 = 4$, igual a la suma de las distancias desde las observaciones mayores que 4, $2 + 2 = 4$.

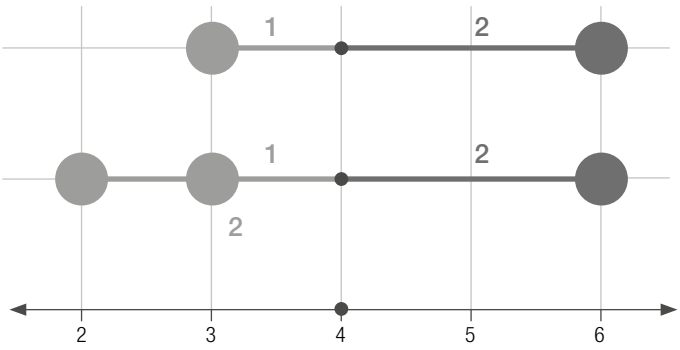


Figura IV.9: La media de las observaciones 2, 3, 3, 6 y 6 es 4.

Supongamos, ahora, que en lugar de una de las personas que asistió a una función de cine **6** veces, se integra al grupo una persona que se desempeña como crítico de cine, que ha asistido **21** veces. Esta situación se ilustra en la **Figura IV.10**, donde se observa que la suma de las distancias desde las observaciones menores que **4, 2 + 1 + 1 = 4**, es bastante menor a la suma de las distancias desde las observaciones mayores que **4, 2 + 17 = 19.**

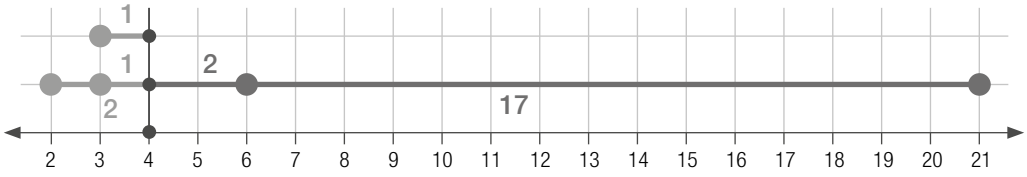


Figura IV.10: Al cambiar el dato de 6 visitas por 21 visitas, la suma de las distancias desde las observaciones mayores que 4 y este, 19, es considerablemente mayor que la suma de las distancias desde las observaciones menores a 4.

Dado que la diferencia entre las sumas es tan grande, debemos mover bastante el valor de la media hacia la derecha, acercándola a la nueva observación, **21**, para disminuir la distancia. La **Figura IV. 11** muestra que, efectivamente, el punto debe moverse hasta el valor **7**, para cumplir la condición de igualdad de sumas de distancias.

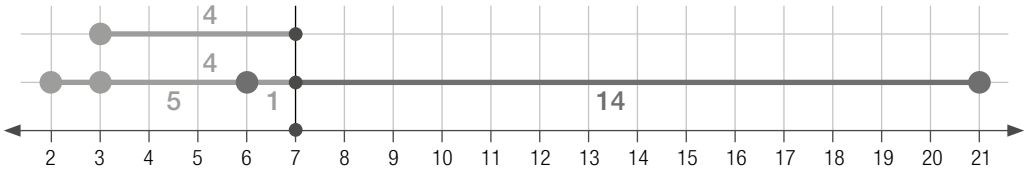


Figura IV.11: La media de las observaciones 2, 3, 3, 6 y 21 es 7.

Notamos entonces, que la presencia de solo una observación extrema, **21** visitas, cambió la media desde **4** visitas al cine en los últimos 6 meses, hasta **7**. De manera figurada, decimos que la observación extrema, de **21** visitas, "atrajo" a la media en 3 unidades, de **4** a **7** visitas.

Por otra parte, veamos qué ocurre con la mediana en esta misma situación. Para ubicar la mediana de los datos originales del grupo, ordenamos las observaciones de menor a mayor valor.

2 3 3 6 6

Y encontramos que la mediana es 3 visitas. Ahora, si reemplazamos una de las personas que asistieron a una función de cine 6 veces durante los últimos 6 meses, por aquel crítico de cine que asistió 21 veces, obtenemos los datos ordenados:

2 3 3 6 21

Donde la mediana no cambia y sigue siendo 3 visitas.

Los ejemplos que hemos dado muestran características de la media y la mediana, como medidas o estadísticos de tendencia central, que son generales como:

- *La media es sensible a valores extremos*: si en el conjunto de datos existe un valor mucho más grande que el grueso de las observaciones, la media crecerá de tal forma que puede no resultar un buen representante del conjunto de ellas. Lo mismo ocurre con un valor mucho más pequeño que el grueso de las observaciones. Es por esto que decimos que los valores extremos "atraen a la media" hacia sí mismos.
- *La mediana no es sensible a valores extremos*: no importa lo grandes o pequeños que sean los valores máximos y mínimos; la mediana no se verá alterada. A modo de ejemplo, en el problema sobre los sueldos de los basquetbolistas, podemos remplazar el sueldo de $4.000.000 por uno aun mayor, como por ejemplo, $10.000.000, y la mediana seguirá siendo la misma, igual a $100.000.

Estas características de la media y la mediana hacen que una comparación entre los valores que ellas toman en un mismo conjunto de observaciones entregue información de la forma de su distribución, esto es, a grandes rasgos, la forma de su histograma. A modo de ejemplo, la **Figura IV.12** presenta 3 situaciones diferentes sobre la distribución de las notas de los alumnos de un curso. En la figura superior, la media y la mediana de las notas son iguales, correspondientes a 4,9, lo que se refleja en que la figura es aproximadamente simétrica en torno a este valor. Por otra parte, la **Figura IV.12**, al centro, muestra otro conjunto de datos, donde, si bien la mayor parte de los alumnos obtuvo notas entre 4,0 y 6,0, dos niños obtuvieron notas considerablemente más bajas, entre 1,0 y 2,0. En este conjunto de datos, la nota media o promedio corresponde a un 4,7, mientras que la mediana corresponde a 5,2. Que la media sea menor a la mediana indica la presencia de algún valor extremo pequeño, como muestran las dos barras de baja altura en el extremo izquierdo del histograma en la **Figura IV.12**, al centro. Finalmente, lo contrario ocurre en la **Figura IV.12**, inferior, donde la distribución muestra barras de baja altura en el extremo derecho de la figura. En este caso, esperaríamos que el valor de la media fuese mayor al de la mediana.

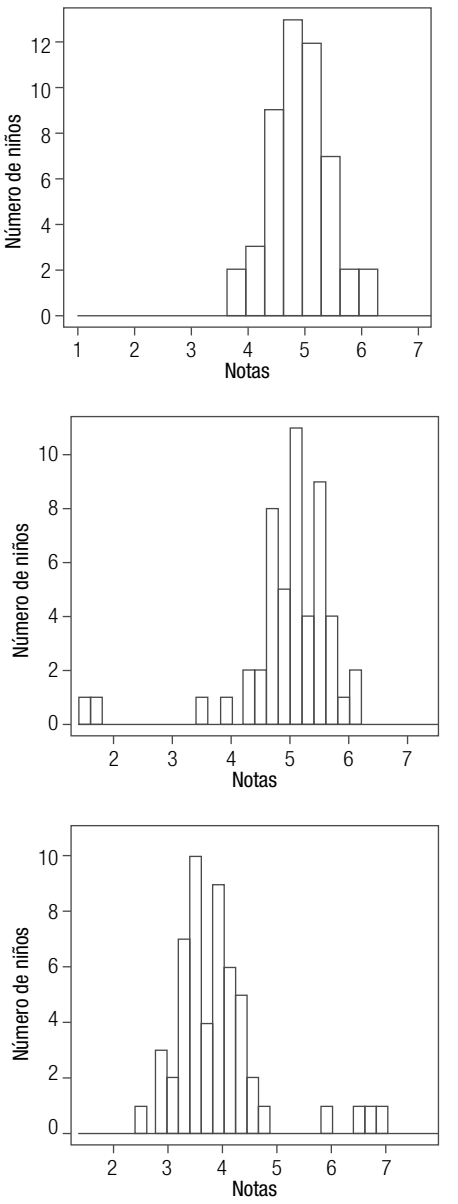


Figura IV.12: Forma de una distribución de acuerdo a la comparación de su media y mediana. Arriba: aproximadamente, iguales media y mediana. Centro: la media es menor que la mediana. Abajo: la media es mayor que la mediana.

En resumen

- La media es sensible a valores extremos o atípicos, muy grandes o muy pequeños con respecto al grueso de los datos. Esto significa que el valor de la media puede cambiar mucho debido a la presencia de este tipo de observaciones.
- Lo anterior no ocurre con la mediana.

Ejercicios

1. En el estudio sobre los efectos del hábito de fumar suele medirse la concentración de cotinina, una sustancia derivada de la nicotina, en la sangre. La siguiente tabla muestra las medias y medianas de las concentraciones de cotinina, en **ng/ml**, en 3 grupos de **100** personas cada uno:

Medida de tendencia central	Fumadores	No fumadores expuestos	No fumadores no expuestos
Media	172,5	60,6	16,4
Mediana	170,0	1,5	0

   a. ¿Cree usted que existe una relación entre el consumo de cigarrillos y la concentración de cotinina en la sangre? Explique.

   b. ¿Por qué cree que la media es bastante mayor que la mediana, tanto para los no fumadores expuestos como para los no expuestos?

2. Una noticia reciente reporta que la encuesta CASEN encontró que el ingreso familiar mensual promedio de nuestro país es de **$800.000**. Emilia había leído en el diario que la mediana del ingreso familiar mensual era de alrededor de **$500.000**, por lo que queda sorprendida por la noticia. ¿Cómo puede usted explicar a Emilia la diferencia entre los dos valores reportados?

3. En un trabajo de verano, se les indica a los postulantes que su salario será variable, aproximadamente **$60.000**. Cuando uno de los postulantes pregunta en qué se basa esta afirmación, el entrevistador le entrega los salarios de los últimos **25** empleados contratados, expresados en pesos:

17.305	478.320	45.678	18.980	17.408	25.676
28.906	12.500	24.540	33.450	12.500	33.855
37.450	20.432	28.956	34.983	36.540	250.921
36.853	16.432	32.654	98.213	48.980	94.024
35.671					

   a. Obtenga la media y la mediana de estos valores. ¿Qué cantidad reportó el entrevistador?

   b. ¿Es el valor reportado, **$60.000**, el valor más representativo de los salarios? ¿en qué sentido?

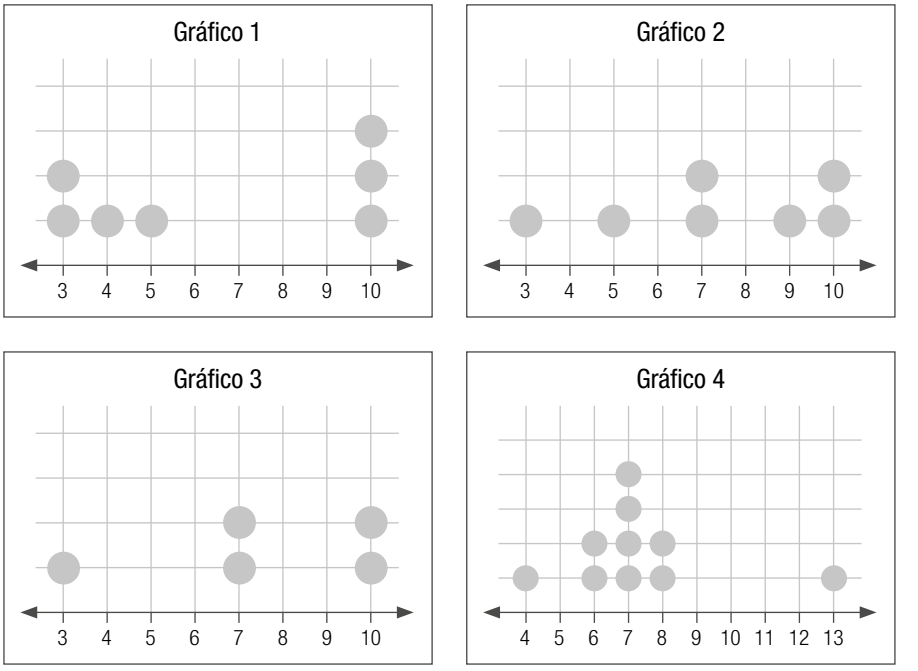
4. Cierta corporación educacional está a cargo de **10** escuelas. Para el próximo año, la corporación ha decidido mantener el número de matrículas en **9** de ellas, y aumentar las matrículas en **600** nuevas vacantes en la escuela ubicada en la zona más populosa.

   a. ¿Cómo afecta el aumento de vacantes en esta última escuela, a la media de los aumentos de vacantes de las escuelas de la corporación?

   b. ¿Cómo afecta el aumento de vacantes en esta última escuela, a la mediana de los aumentos de vacantes de las escuelas de la corporación?

c. ¿Qué medida de tendencia central, media o mediana, utilizaría como medida de resumen de los aumentos de vacantes en las escuelas de la corporación? Explique de qué puede depender esta decisión.

5. Observe los siguientes gráficos de puntos.



a. En cada caso, entregue valores aproximados de media y mediana, sin efectuar cálculos numéricos ni procedimientos habituales.

b. ¿En qué conjunto de datos cree usted que la media se ve afectada por valores extremos? ¿de qué forma?

c. Realizando ahora los cálculos numéricos habituales, ¿cuánto afecta el(los) valor(es) extremo(s) a la media en los gráficos 3 y 4? Realizando el procedimiento habitual, ¿se ve afectada la mediana?

2.4. La moda

Retomemos el problema del inicio del **Capítulo 3**, donde se pregunta a un grupo de 15 niños sobre sus deportes favoritos. Replicamos en la **Tabla IV.2** las frecuencias absolutas obtenidas en base a sus respuestas.

Deporte favorito	Número de alumnos
Fútbol	6
Básquetbol	4
Gimnasia rítmica	3
Tenis	2

Tabla IV.2: Deportes favoritos de los niños de un curso.

En base a la tabla, podemos decir que la mayoría [1] de los niños prefiere el fútbol, o que el fútbol es el deporte favorito que más se repite. En ese sentido, podemos decir que el fútbol es el deporte más popular. Basado en esta idea, se dice que, en este caso, el fútbol corresponde a la moda de la distribución de los deportes favoritos de los niños.

También podemos extraer la moda a partir del gráfico de barras que representa los datos de la **Tabla IV.2**. En efecto, en la **Figura IV.13** se observa que la categoría fútbol tiene asociada la barra de mayor altura, lo que indica que su frecuencia es mayor que la de los deportes restantes. Notemos que también sería posible extraer esta información a partir de un gráfico de barras que represente frecuencias relativas o relativas porcentuales, en lugar de frecuencias absolutas, como la **Figura IV.13**.

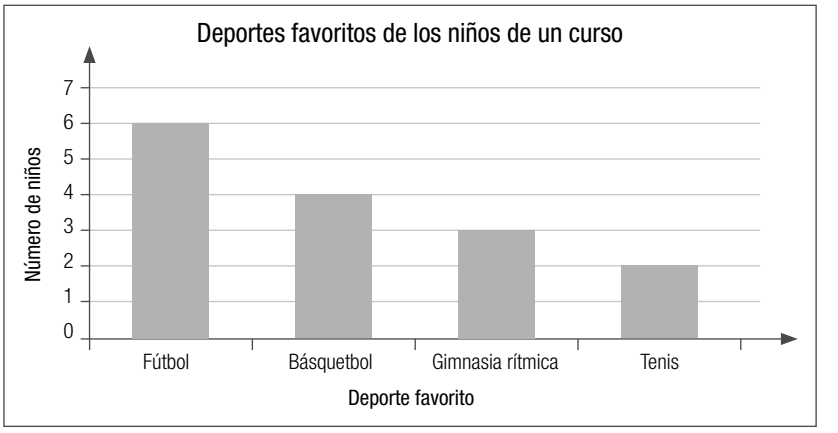


Figura IV.13: Deportes favoritos de los niños. El fútbol corresponde a la moda.

En otro ejemplo, la **Figura IV.14** muestra dos gráficos de barras de datos ficticios sobre colores favoritos. En ambos gráficos, la moda corresponde al color verde, dado que las barras de mayor altura se encuentran sobre este color, sin embargo, en el gráfico de la izquierda, el verde es un mejor representante de las observaciones que en el gráfico de la derecha. Esto se debe a que en el gráfico de la izquierda, la frecuencia del color verde es considerablemente mayor que las frecuencias de

[1] Entendemos mayoría como el grupo más grande de observaciones que comparte su valor, a diferencia del grupo formado por más de la mitad de las observaciones que comparten su valor.

los colores restantes. En la figura de la derecha, aunque el color verde es el de mayor frecuencia, esta es muy similar a las frecuencias de los colores restantes, lo que hace que la moda no sea tan buen representante de las observaciones en este caso. Vemos que reportar únicamente la moda diciendo "el color más popular es el verde" no nos entrega información acerca de su calidad como representante, lo cual puede ser considerado como una debilidad de la moda.

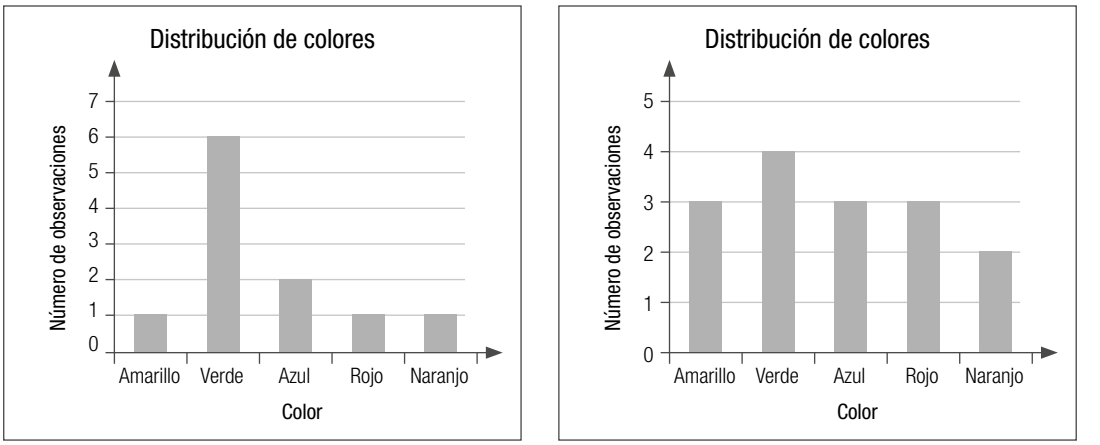


Figura IV.14: En ambas figuras, la moda corresponde al color verde. Sin embargo, este color representa de mejor manera a las observaciones en el gráfico de la izquierda.

Notemos que, aunque la moda sea referida como una medida de tendencia central, esta no responde en realidad al concepto de centro. A modo de ejemplo, en la **Figura IV.14**, por tratarse de una variable cualitativa nominal, podríamos reordenar las categorías en el eje horizontal, ubicando el color verde en uno de los extremos y no necesariamente al centro.

En ambos ejemplos anteriores, hemos elegido variables cualitativas, como deporte favorito o color. Veamos ahora qué ocurre cuando la variable de interés corresponde a una variable cuantitativa, partiendo por el caso de una variable discreta que toma pocos valores.

A modo de ejemplo, la **Figura IV.15** muestra el histograma correspondiente al número de horas que dedicó cada uno de **12** niños a practicar deportes durante la semana anterior. En la figura podemos identificar que la moda corresponde a **3** horas, puesto que tiene asociada la barra de mayor altura.

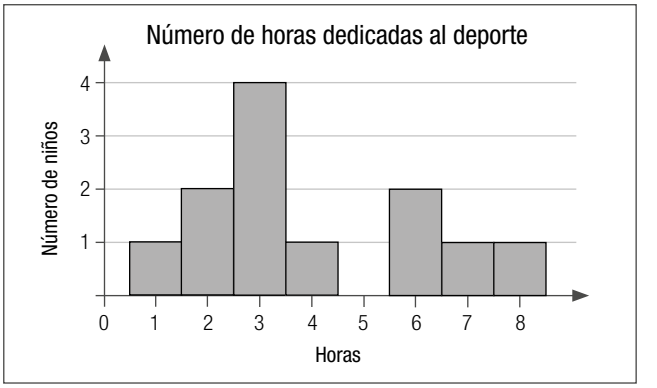


Figura IV.15: Número de horas dedicadas por cada uno de 12 niños a practicar deportes durante la semana anterior. La moda corresponde a 3 horas.

La situación es diferente si consideramos variables cuantitativas continuas o discretas que toman un gran número de valores. En efecto, consideremos las siguientes observaciones, que corresponden a los pesos, en gramos, de chocolates con almendras:

0,957	0,912	0,915	0,925	0,912	0,911
0,913	0,958	0,947	0,920	0,886	0,914

Para estos datos, la moda corresponde a **0,912** gramos, valor que se observa **2** veces.

Para pensar

¿Cree usted que este valor es, de alguna manera, un buen representante de los pesos de los chocolates del conjunto de datos? ¿a qué puede deberse?

En general, para variables cuantitativas, tanto continuas como discretas que tomen muchos valores diferentes, es altamente improbable que se observe un mismo valor de la variable más de una vez y, si se observan, ellos probablemente presentan frecuencias muy pequeñas que no los hacen más representativos que el resto de las observaciones. Una opción en estas situaciones corresponde a identificar una *clase modal*. Consideremos por ejemplo, el histograma de la edad de los trabajadores en una empresa, medida en años, que se muestra en la **Figura IV.16**. Observamos que la barra más alta corresponde a la categoría entre **30** y **35** años. Podemos decir que esta categoría o clase corresponde a la clase modal, pues es la categoría que contiene el mayor número de observaciones.

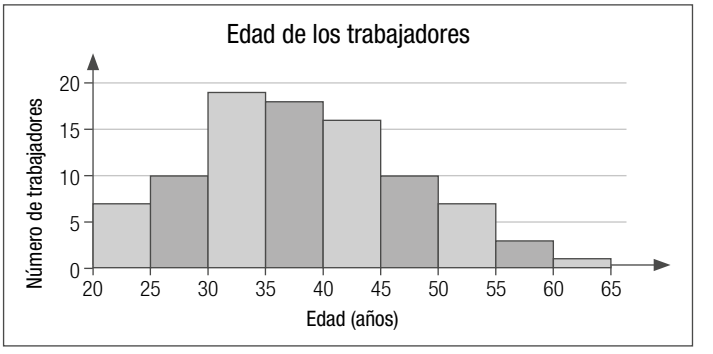


Figura IV.16: Edad de los trabajadores de una empresa.

Sin embargo, sabemos que no hay una única manera de determinar el número y la amplitud de las clases a considerar en la construcción de un histograma. Para cada definición de intervalos de clase, se encontrará una clase modal diferente, lo que representa un inconveniente de esta estrategia. Debido a esto, no es recomendable utilizar la moda como una medida de resumen de las observaciones para variables cuantitativas continuas, o discretas que tomen muchos valores diferentes.

Al igual como ocurre con variables cualitativas, notemos nuevamente que, en el histograma de las edades de los trabajadores en la **Figura IV.16**, si bien la clase modal corresponde al grupo de trabajadores entre **30** y **35** años, el número de observaciones es bastante similar en la categoría entre **35** y **40** años, **19** y **18** trabajadores en cada categoría, respectivamente. Nuevamente, reportar únicamente la clase modal no refleja la calidad de esta como representante de las observaciones.

En el ejemplo sobre el número de horas que 12 niños dedicaron a practicar un deporte durante la semana anterior, que corresponde a una variable cuantitativa que toma pocos valores, también es posible ver que la moda no responde a un concepto de centro. En efecto, si la mayoría de los niños, por ejemplo, 7 de ellos, no hubiese practicado un deporte durante la semana anterior, la moda correspondería a la barra ubicada sobre el valor 0, es decir, en el extremo izquierdo, y no correspondería a una noción de centro.

Consideremos, finalmente, la Figura IV.17, que muestra las frecuencias absolutas de las mascotas de un grupo de niños. Tanto perros como gatos son mascotas igualmente populares, y son ambas las mascotas de 6 niños. En este caso, la moda no es única y decimos que la distribución es bimodal. Pueden también existir distribuciones trimodales, con tres modas, o polimodales, en general.

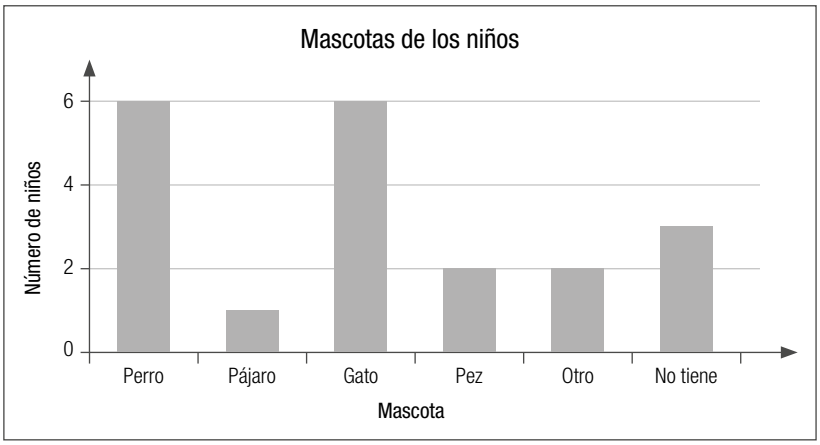


Figura IV.17: Mascotas de los niños de un curso.

En resumen

- La *moda* corresponde al valor que más se repite en el conjunto de datos.
- No tiene sentido reportar la moda cuando las observaciones son cuantitativas continuas, o cuantitativas discretas tomando muchos valores diferentes.
- Una desventaja de la moda es que el conocer únicamente su valor no nos indica su calidad como representante: aunque sea el valor que más se repite, esto puede ocurrir un número muy pequeño de veces o pueden existir otros valores que se repitan un número de veces que, aunque sea menor, sea muy similar al número de veces que se repite la moda.
- Si bien la moda es considerada una medida de tendencia central, no responde a un concepto de centro.
- La moda de un conjunto de datos puede no ser única.

2.5. Errores y dificultades relacionadas a medidas de tendencia central

Algunos errores relacionados al cálculo e interpretación de la media, se deben principalmente a dificultades de comprensión del concepto.

Algunos errores que se presentan en el cálculo de la media son:

- *Confundir las frecuencias de las observaciones con los valores que estas toman, cuando los datos se presentan de manera agrupada en una tabla de frecuencias*. A modo de ejemplo, consideremos la **Tabla IV.3**, que muestra las tallas de zapatos de los niños de un curso.

Talla de zapatos	30	31	32	33	34
Número de niños	2	4	6	2	1

Tabla IV.3: Talla de zapatos que calzan los niños de un curso.

En un caso como este, un cálculo equivocado de la media corresponde a:

$$\frac{2+4+6+2+1}{5}=3$$

Ya que se han confundido las frecuencias, el número de niños, con los valores que toma la variable de interés, talla de zapatos. En efecto, los valores que esta última toma en el conjunto de datos de la tabla son, en realidad, **30**, **31**, **32**, **33** y **34**.

- En la misma situación, otro error corresponde a *olvidar que cada uno de los valores observados, que se entregan en la tabla, puede haber ocurrido más de una vez*, según indique su frecuencia. De este modo, otro cálculo equivocado de la media corresponde a:

$$\frac{30+31+32+33+34}{5}=32$$

Donde se ha olvidado ponderar el valor de la observación por su frecuencia o número de veces que se repite. En efecto, el cálculo correcto de la media corresponde a:

$$\frac{30\cdot 2+31\cdot 4+32\cdot 6+33\cdot 2+34\cdot 1}{2+4+6+2+1}=\frac{476}{15}$$

O, aproximadamente, **31,73**.

Un error similar suele presentarse al obtener la media de un conjunto de datos a partir de un gráfico de puntos como el que se muestra en la **Figura IV.18**. En situaciones como estas, no se debe olvidar repetir cada uno de los valores de las observaciones, en este caso, **30**, **31**, **32**, **33** y **34**, el número de veces que se observan, lo que se puede obtener contando los puntos en cada uno de ellos, a partir de la figura.

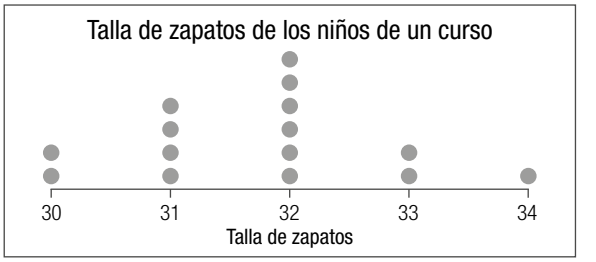


Figura IV.18: Talla de zapatos de los niños de un curso.

- *Ignorar observaciones cuyo valor es cero*. A modo de ejemplo, si el conjunto de observaciones corresponde a **5**, **5**, **4**, **2**, **3**, **0** y **1**, un cálculo erróneo de la media corresponde a:

$$\frac{5 + 5 + 4 + 2 + 3 + 1}{6}$$

Donde se ignora la observación de valor cero al determinar el número de observaciones en el denominador, y se divide entonces por **6** y no por el verdadero número, **7**.

- *No ponderar las medias de subgrupos del conjunto de datos, cuando estos contienen diferentes números de observaciones*. A modo de ejemplo, considere el siguiente problema:

"En un ascensor hay 10 personas, de las cuales 4 son mujeres y 6 son hombres.
La media del peso de las mujeres es **60** kilos y la de los hombres es de **80** kilos.
¿Cuál es la media del peso de las 10 personas?"

Podríamos, erróneamente, obtener la media como:

$$\frac{60 + 80}{2} = 70 \text{ kilos}$$

Olvidando que los grupos de hombres y mujeres son de diferentes tamaños, por lo que no debieran tener la misma ponderación en el cálculo. En efecto, dado que hay 4 mujeres y 6 hombres, la media puede obtenerse de manera correcta como:

$$\frac{4 \cdot 60 + 6 \cdot 80}{10} = 72 \text{ kilos}$$

Por otra parte, algunos de los errores que se presentan en la interpretación de la media pueden ser:

- *Confundir el concepto de media con los conceptos de mediana o moda*. En efecto, usos coloquiales de la media suelen llevar a su malinterpretación, entendiendo que el valor reportado corresponde al valor más frecuente, que corresponde a la moda, o al punto medio, que corresponde a la mediana.

- *Determinar la media erróneamente a partir de un histograma*, como el que se muestra en la **Figura IV.19**. En casos como este, suele identificarse la media como un valor muy frecuente. A modo de ejemplo, en la figura, la barra más alta se encuentra cercana a un **5,0**. Sin embargo, eso no significa que la media sea cercana a este valor.

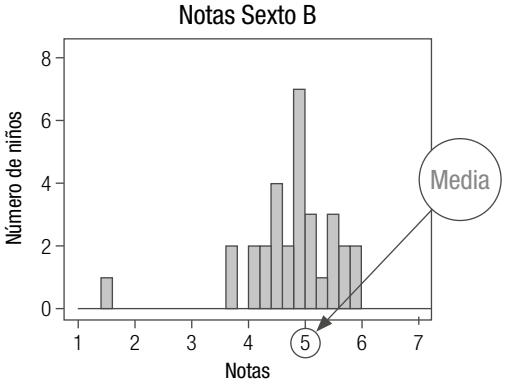


Figura IV.19: Equivocadamente, suele identificarse la media como el valor más frecuente.

- *Obtener la media cuando se trata de variables cualitativas*. En particular, un error corresponde a obtener la media de observaciones cualitativas ordinales, medidas en una escala numérica. A modo de ejemplo, una escala de acuerdo-desacuerdo puede corresponder a: Totalmente en desacuerdo - En desacuerdo - Indiferente - De acuerdo - Totalmente de acuerdo, asignándose a estas categorías los valores 1, 2, 3, 4 y 5, respectivamente. En este caso, no es razonable trabajar con la media, puesto que, dado que las etiquetas utilizadas, 1, 2, 3, 4 y 5, están todas a la misma distancia, se estaría asumiendo que las categorías asociadas, es decir, los grados de acuerdo-desacuerdo, también están a la misma distancia unos de otros. Sin embargo, esto resulta artificial, por tratarse de actitudes no medibles numéricamente.
- *Redondear la media por pensar que debe corresponder a uno de los valores de la variable de interés*. A modo de ejemplo, supongamos que la media del número de hermanos de un grupo de alumnos es 2,4 hermanos. Un error corresponde a redondear este valor, debido a que no corresponde a un número de hermanos válido, y reportar una media de 2 hermanos, en lugar de 2,4, que es lo correcto.

Por otra parte, algunos de los errores relacionados a la mediana suelen ser:

- *Obtener la mediana como el valor central de las observaciones, o el promedio de los dos valores centrales, sin haber ordenado previamente los datos según su magnitud*. Esto muestra un foco en aspectos procedimentales, y no realmente en el concepto de mediana.
- *Obtener la mediana cuando se trata de variables cualitativas ordinales.*

La siguiente lista de ejercicios se relaciona a las dificultades y errores a los que nos hemos referido.

Ejercicios

1. La profesora Palma y su curso están resolviendo problemas que requieren determinar la media de un conjunto de datos. La profesora presenta el siguiente problema:

   "En una sala de clases hay 15 niñas y 11 niños. Si el promedio de peso de niñas y niños juntos es 55 kilos y el promedio de peso de las niñas es 48, ¿cuál es el promedio de peso de los niños?".

   La profesora observa que los niños llevan a cabo cálculos, correctos e incorrectos, para determinar la respuesta.

   a. Describa dos cálculos incorrectos que pudieron haber realizado los alumnos. Indique qué concepción errónea de la media puede haberlos causado y por qué son incorrectos.

   b. Describa dos cálculos que permitan obtener la respuesta correcta.

2. Suponga que usted desea utilizar el siguiente enunciado para crear una pregunta de selección múltiple:

   "La municipalidad de cierta comuna en el sur de Chile desea determinar la media del número de niños por familia. Para esto, el encargado divide el número total de niños de la ciudad por 50, que es el número de familias. ¿Cuál de las siguientes frases debe ser cierta, si la media del número de niños por familia es 2,2?".

   Construya una respuesta correcta y cuatro distractores, uno de los cuales detecte si el alumno, erróneamente, entiende que el valor de la mediana debe ser cercano al valor de la media.

3. Para cada una de las dos situaciones que se presentan, proponga un problema de selección múltiple, con 3 distractores, que cubran los errores mencionados.

   a. Presentar la moda como preferible a la media, por ser mejor representante de los datos.

   b. Proponer descartar valores nulos.

4. Un profesor utilizó el siguiente ejercicio en una evaluación:

   "La mediana del siguiente conjunto de números: 1, 5, 1, 6, 1, 6 y 8, es:

   a. 1 b. 4 c. 5 d. 6"

   Para cada uno de los distractores, identifique el error que el profesor intenta detectar con él.

La siguiente lista de ejercicios integra lo que hemos estudiado hasta ahora, referente a medidas de tendencia central: media, mediana y moda.

Ejercicios

1. Considere el siguiente conjunto de datos:

Valor	1	2	3	4	5	6	7	8	9
Frecuencia absoluta	9	8	7	6	5	6	7	8	9

   Para obtener la media a partir de esta tabla de frecuencias podemos realizar el cálculo:

$$\frac{9 \cdot 1 + 8 \cdot 2 + 7 \cdot 3 + 6 \cdot 4 + 5 \cdot 5 + 6 \cdot 6 + 7 \cdot 7 + 8 \cdot 8 + 9 \cdot 9}{9 + 8 + 7 + 6 + 5 + 6 + 7 + 8 + 9} = \frac{325}{65} = 5$$

   Luego, la media de los datos es 5.

   Para la mediana, podemos obtener la tabla de frecuencias absolutas acumuladas:

Valor	1	2	3	4	5	6	7	8	9
Frecuencia absoluta acumulada	9	17	24	30	35	41	48	56	65

   Vemos que se tiene un total de 65 observaciones, por lo que la mediana corresponde al valor de la observación ubicada en el lugar 33, al ordenarlas de menor a mayor valor. El primer valor cuya frecuencia absoluta acumulada es mayor o igual a 33 corresponde a 5 (su frecuencia absoluta acumulada es 35). Luego, la mediana corresponde a 5.

   Podemos obtener la moda a partir de la tabla original, donde hay 2 observaciones que más se repiten: 1 y 9 (9 veces cada una). Luego, la distribución de los datos es bimodal y sus modas son 1 y 9.

De estos 3 valores, los más representativos son la media o la mediana, dado que las modas se repiten solo una vez más que las observaciones que le siguen (2 y 8 se repiten 8 veces cada una).

a. Repita los procedimientos anteriores para obtener media, mediana y moda en el siguiente conjunto de datos. ¿Cuál(es) de estos 3 valores considera más representativo(s) en este caso?

Valor	1	2	3	4	5	6	7	8	9
Frecuencia absoluta	7	20	15	11	8	3	2	0	15

b. Repita los procedimientos anteriores para obtener media, mediana y moda en el siguiente conjunto de datos. ¿Cuál(es) de estos tres valores considera más representativo(s) en este caso?

Valor	1	2	3	4	5	6	7	8	9
Frecuencia absoluta	6	1	2	3	5	5	4	3	0

2. La siguiente tabla entrega las notas obtenidas por los alumnos de un curso en una prueba.

Nota	Número de alumnos
6,3	2
6,4	1
6,5	6
6,6	5
6,7	13
6,8	4

Obtenga la media, la mediana y la moda de estas notas y comente relaciones observadas entre ellas.

3. Los siguientes son los puntajes obtenidos por 20 pacientes en un test de esfuerzo:

93 84 97 98 100 78 86 100 85 92
72 55 91 90 75 94 83 60 81 95

a. Encuentre la mediana y la moda de estos puntajes.

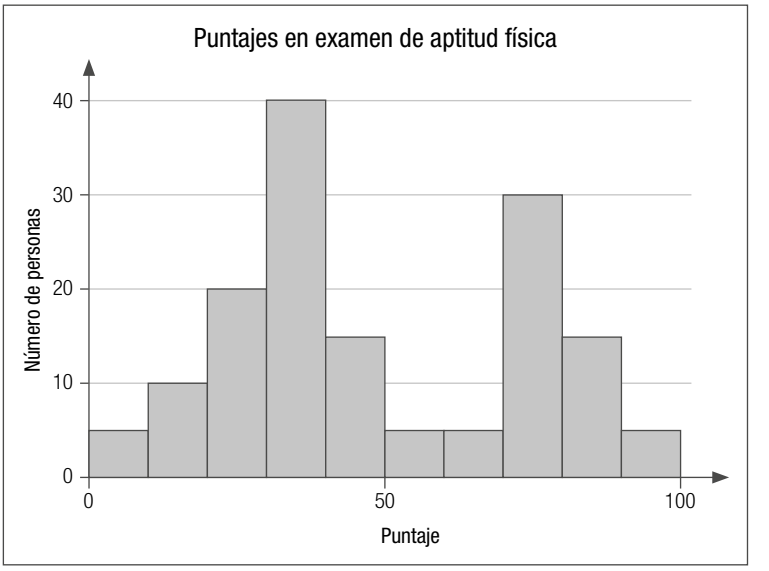
b. Complete la siguiente tabla de frecuencias y encuentre la clase que contiene a la mediana, y la clase modal. Compare con sus respuestas en a.

Porcentaje	Número de pacientes
51 a 60	
61 a 70	
71 a 80	
81 a 90	
91 a 100	
Total	20

4. En una escuela, se venden cada día 5 alternativas de postre. El dueño del casino registró el número de cada uno de los 5 tipos que vendió en una semana en particular. ¿Qué medida de tendencia central debiese utilizar para decidir cuál es el postre favorito de los niños?

5. Un programa computacional reportó las notas de los niños de un curso y los resultados fueron muy buenos. De hecho, todos los niños obtuvieron una nota superior a 5,5. Para conocer el rendimiento global del curso, la profesora pide al programa que le entregue las medidas de tendencia central de las notas: media, mediana y moda. Sin embargo, ella debe tener en cuenta que las notas de 4 niños que faltaron a la prueba aparecen registradas como 1,0. ¿Qué medida de tendencia central será preferible utilizar como representante en este caso? Justifique.

6. El histograma en la figura muestra la distribución de los puntajes de logro en un examen de aptitud física.



   a. ¿Qué medida de tendencia central, media, mediana o moda, cree usted que sería más representativa en un caso como este? Explique.
   b. Entregue una estimación de media, de mediana y de moda basadas en la figura.
   c. ¿Puede obtener estas medidas de manera exacta?

3. Medidas de posición relativa

Como vimos anteriormente, una vez que las observaciones han sido ordenadas según su magnitud, la mediana las divide en 2 grupos que contienen el mismo número de observaciones. Por otra parte, también vimos que, cuando pocas observaciones toman el valor de la mediana en relación al total de observaciones en el conjunto de datos, también puede decirse que, aproximadamente, el 50% de las observaciones es menor o igual a la mediana y, aproximadamente, el 50% es mayor o igual a ella. Argumentamos, por ejemplo, que una situación de este tipo ocurre en conjuntos de datos como los puntajes del SIMCE de todos los alumnos del país. La característica de conjuntos de datos de este tipo es que la variable de interés puede tomar muchos valores diferentes, y el número de observaciones es suficientemente grande, de modo que la proporción de valores que se repiten es pequeña, en comparación a este número.

Según veremos, las medidas de posición relativa corresponden, en este sentido, a una extensión del concepto de mediana y, al igual que esta, ayudan a describir la distribución de los datos. De manera intuitiva, aunque luego daremos una definición formal, las medidas de posición relativa son valores que intentan dividir a las observaciones, una vez ordenadas según magnitud creciente, en un número de subgrupos dado, de modo que todos ellos contengan el mismo número de observaciones.

Las medidas de posición relativa que estudiaremos corresponden a *cuartiles, quintiles, deciles* y *percentiles* que, por ahora, diremos que dividen al conjunto de observaciones en 4, 5, 10 y 100 grupos del mismo tamaño, respectivamente. Veremos, sin embargo, que esto no es siempre posible.

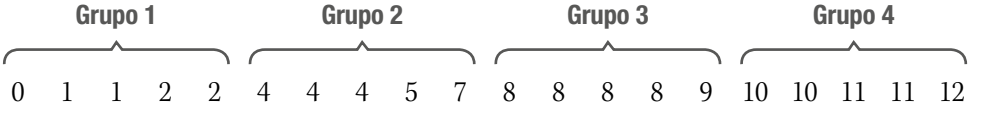
Si bien daremos ejemplos sobre la manera de obtener las medidas de posición relativa de interés, el énfasis estará puesto en su interpretación, y nos concentraremos en los casos en que la variable toma muchos valores y el tamaño del conjunto de datos es grande.

3.1. Cuartiles

Consideremos las siguientes observaciones, que corresponden al número de horas, aproximado, que cada uno de los niños de un grupo dedicó a leer durante la semana pasada, ordenados de menor a mayor:

0 1 1 2 2 4 4 4 5 7 8 8 8
8 9 10 10 11 11 12

Quisiéramos resumir estas observaciones describiendo su distribución, lo que haremos a partir de la interpretación de sus *cuartiles*, los que, inicialmente, supondremos que dividen al conjunto de observaciones en 4 grupos de igual tamaño, manteniendo el ordenamiento mostrado. Dado que en total hay 20 observaciones, cada uno de los grupos generados debe contener 5 observaciones. Así, los 4 grupos son:



Los cuartiles corresponden a valores que separan un grupo de datos del siguiente. Consideremos los Grupos 1 y 2. Al igual como ocurre con la mediana cuando el número de observaciones es par, en este caso también existen infinitos valores que pueden separar estos 2 grupos, dado que cualquier valor entre 2 y 4 horas servirá para esto. Podemos tomar la convención de elegir como representante al punto medio entre estos dos valores, es decir $\frac{(2+4)}{2} = 3$ horas. Del mismo modo, cualquier valor entre 7 y 8 horas separa los Grupos 2 y 3, por lo que, según la convención que hemos tomado, elegimos 7,5 horas. Finalmente, elegimos el valor 9,5 horas como valor que separa los Grupos 3 y 4.

Notamos que para separar el conjunto de observaciones en 4 grupos hemos requerido de 3 puntos. Cada uno de ellos se denomina *cuartil*. Para identificarlos, hablamos del primer, segundo y tercer cuartil. De este modo, en el conjunto de datos, el primer cuartil, que se anota como Q_1, corresponde a 3 horas, el segundo cuartil, que se anota como Q_2, corresponde a 7,5 horas, y el tercer cuartil, que se anota como Q_3, corresponde a 9,5 horas [2]. Existe, además, la convención de numerar los cuartiles desde los valores pequeños a los más grandes, por lo que, a diferencia de la mediana, no da lo mismo si ordenamos las observaciones de manera creciente o decreciente. Como hemos hecho hasta ahora, en todo lo que sigue asumiremos que las observaciones se encuentran ordenadas de manera creciente.

Para pensar

¿En cuántos grupos de igual número de observaciones divide el segundo cuartil al conjunto de datos? Explique la relación entre el segundo cuartil y la mediana.

La **Figura IV.20** muestra una manera de representar gráficamente los cuartiles. En ella, cada cuadrado representa una observación o un niño. De este modo, la figura muestra 20 cuadrados, dado que hay 20 niños. A modo de ejemplo, como 2 niños dedicaron 1 hora a leer durante la semana pasada, dibujamos 2 cuadrados, uno sobre el otro, en el valor 1 del eje de las abscisas. Dado que ningún niño dedicó exactamente 6 horas a leer durante la semana pasada, no dibujamos cuadrados en el valor 6 del mismo eje. La figura muestra, a través de líneas punteadas, los valores de los cuartiles que obtuvimos anteriormente, 3, 7,5 y 9,5 horas, e indica, a través de diferentes colores, los 4 grupos de observaciones que ellos determinan.

[2] La notación Q_1, Q_2 y Q_3 hace alusión a la denominación de los cuartiles en inglés, *quartile*.

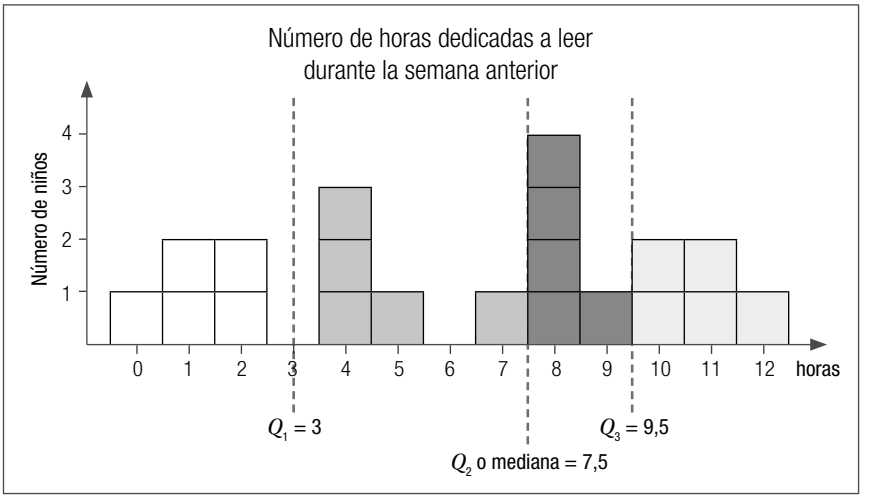


Figura IV.20: Representación de los cuartiles. Cada uno de los 20 niños se representa a través de un cuadrado ubicado en el número de horas que dedicó a leer durante la semana pasada.

La **Figura IV.21** corresponde a la **Figura IV.20** modificada, donde los cuadrados han sido integrados en una barra, obteniéndose el histograma de los datos.

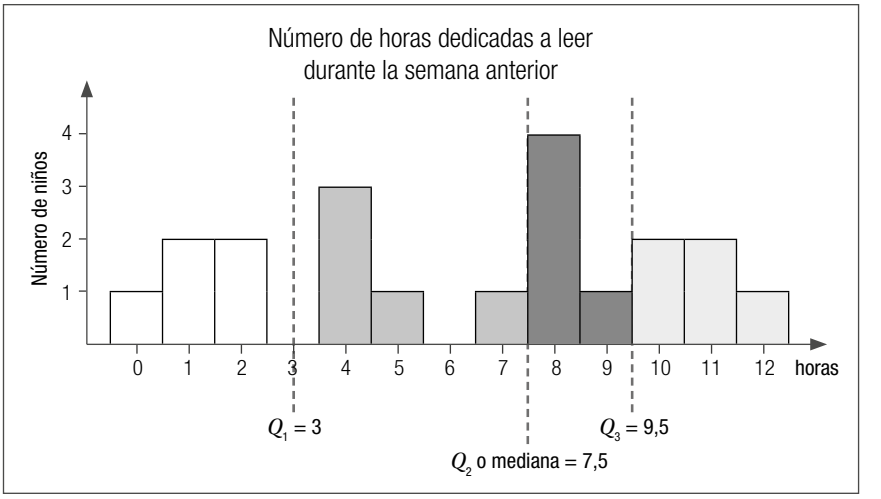


Figura IV.21: Histograma de los datos. Se han utilizado barras de diferentes colores para identificar las observaciones en los diferentes cuartiles.

Las Figuras **IV.20** y **IV.21** resaltan las principales características de los cuartiles:

- En la Figura **IV.20**, dado que hay un total de **20** niños, el número de niños, o cuadrados, en cada uno de los cuatro grupos que determinan los cuartiles es siempre $\frac{20}{4} = 5$.
- En la Figura **IV.21**, dado que el área de cada barra de un histograma es proporcional a la frecuencia (absoluta o relativa) de la categoría que esta representa, cada una de las 4 áreas determinadas por los cuartiles, y diferenciadas por un color, representa un $\frac{100}{4}\% = 25\%$ del área total de las barras.

- En la Figura IV. 20, el 50% de las observaciones, es decir, 10, es menor que el segundo cuartil. Esto significa que el segundo cuartil corresponde a la mediana.
- Las distancias entre los cuartiles no son las mismas. En efecto, la distancia entre el primer y segundo cuartil o mediana, es 7,5 – 3 = 4,5 horas, mientras que la distancia entre el segundo cuartil o mediana, y el tercero, es de 9,7 – 7,5 = 2 horas. Aunque, en casos particulares, es posible que las distancias entre los cuartiles sean las mismas, no es una característica necesaria de estos, así como de ninguna medida de posición relativa.

Para entender por qué las distancias entre los cuartiles son, en general, diferentes, notemos que las observaciones pueden estar más concentradas en algunos valores que en otros. A modo de ejemplo, existen 4 niños que leyeron 8 horas durante la semana pasada, lo que concentra observaciones en ese valor, pero no hay niños que leyeran 3 o 6 horas. Así, por ejemplo, cuando nos movemos desde la mediana hacia la derecha para abarcar un nuevo 25% de los datos que, en este caso, corresponde a 5 observaciones, nos encontramos con que al movernos únicamente 0,5 horas ya capturamos 4 de ellas. Para alcanzar la quinta, basta moverse una unidad a la derecha, es decir, a 9 horas. De este modo, la distancia entre la mediana y el tercer cuartil es relativamente pequeña, en comparación con la distancia entre el primer cuartil y la mediana. En este último caso, debemos movernos un mayor número de unidades, u horas, hacia la derecha para abarcar las 5 observaciones requeridas. Esto nos sugiere que si 2 cuartiles se encuentran muy distantes, entonces, las observaciones están muy dispersas en dicho intervalo.

Una vez obtenidos los cuartiles de la distribución del número de horas que dedicó cada niño a leer durante la semana pasada, podemos utilizarlos para dar la descripción deseada. Podemos afirmar que: dado que el primer cuartil es 3 horas, un 25% de los niños dedicó menos de 3 horas a leer durante la semana pasada. El segundo cuartil, 7,5 horas, indica que un 50% de los niños dedicó a leer menos de 7,5 horas durante la semana pasada, mientras que el tercer cuartil, 9,5 horas indica que un 75% de los niños dedicó a leer menos de 9,5 horas la semana pasada.

También podríamos decir que un 25% de los niños dedicó entre 3 y 7,5 horas a leer, o que otro 25% de los niños dedicó entre 7,5 y 9,5 horas a leer, o que un 25% de los niños dedicó más de 9,5 horas a leer.

3.2 Otras interpretaciones de los cuartiles como medidas de posición relativa

Retomemos el ejemplo en las Figuras IV.20 y IV.21. Si bien en dicho ejemplo hemos elegido el número de observaciones y los valores que estas toman de modo de identificar claramente los cuartiles, no siempre es posible dividir el conjunto de datos en 4 grupos de exactamente igual tamaño, así como tampoco que las divisiones estén dadas exactamente entre 2 barras de un histograma, como en la Figura IV.21.

Esta inquietud no solo es válida cuando se desea encontrar los cuartiles de un conjunto de observaciones, sino para cualquier medida de posición relativa. La importancia de discutir este punto radica en la forma en que debemos interpretar los valores obtenidos.

Tal como anticipamos, no solo para los cuartiles, sino para cualquier medida de posición relativa, no siempre es posible encontrar puntos tales que los grupos deseados tengan exactamente el mismo tamaño. Esto no significa que las medidas de posición relativa de interés no existan, sino que debemos considerar una definición más amplia de estas para encontrarlas. En general, para cualquier valor q entre 0 y 100, que representará un porcentaje de las observaciones, se denomina *cuantil* q a un valor tal que:

i. Al menos un q% de las observaciones son menores o iguales a ella.

ii. Al menos un $(100 - q)$% de las observaciones son mayores o iguales a ella.

Explicaremos esto a través del ejemplo en las **Figuras IV.20** y **IV.21**, tomando $q = 25$. Notemos que:

- 5 observaciones, es decir, un $\frac{5}{20} = 25\%$ de las observaciones, son menores que el primer cuartil, 3 horas.
- 15 observaciones, es decir, un $\frac{15}{20} = 75\%$ de las observaciones, son mayores que el primer cuartil, 3 horas.

Es decir, el primer cuartil cumple con las 2 condiciones pedidas, i. e ii., con $q = 25$. Del mismo modo, encontraremos, por ejemplo, que la mediana corresponde al caso $q = 50$, y el tercer cuartil al caso $q = 75$.

Podríamos, entonces, utilizar las dos condiciones, i. e ii., para encontrar los cuartiles deseados en el caso general, por ejemplo, cuando el número de observaciones no es divisible por 4. Para ilustrar, supongamos un conjunto de 33 observaciones y que estas se encuentran previamente ordenadas de manera creciente. Para facilitar la compresión, asumiremos que no existen valores de las observaciones repetidos. Notemos que el número de observaciones no permite dividirlas en 4 grupos de exactamente el mismo tamaño, puesto que 33 no es divisible por 4. Para encontrar el primer cuartil, consideremos entonces $q = 25$. Debemos encontrar una observación tal que:

- Al menos un 25% de las observaciones son menores o iguales a ella.
- Al menos un $(100 - 25)\% = 75\%$ de las observaciones son mayores o iguales a ella.

Consideremos la observación que, al estar todas ordenadas en orden creciente de magnitud, se encuentra en la novena posición. El número de observaciones menores o iguales a ella es 9, es decir, $\frac{9}{33}$ del total de las observaciones o, aproximadamente, un 27,3%. Es decir, la novena observación cumple con la primera propiedad: al menos un 25% de las observaciones son menores o iguales a ella. Por otra parte, el número de observaciones mayores o iguales a ella es $(33 - 8) = 25$, es decir, $\frac{25}{33}$ del total de las observaciones o, aproximadamente, un 75,8%. Es decir, la novena observación cumple con la segunda propiedad: al menos un $(100 - 25)\% = 75\%$ de las observaciones son mayores o iguales a ella. Luego, la observación en la novena posición es candidata a primer cuartil.

Notemos que las observaciones vecinas, a la izquierda y la derecha de la novena observación, es decir, las observaciones 8 y 10, no cumplen con las propiedades de interés. En efecto, para la observación 8, encontraremos que los porcentajes de interés son, aproximadamente, 24,2% y 78,8%. Entonces, esta observación no cumple con la primera propiedad. Para la observación 10, encontraremos que los porcentajes de interés corresponden a, aproximadamente, 30,3% y 72,7%, es decir, esta observación no cumple con la segunda propiedad. Concluimos, entonces, que podemos tomar la novena observación, al ordenarlas en orden de magnitud creciente, como el primer cuartil.

En el caso en que más de una observación del conjunto de datos cumpla con las condiciones pedidas, se considera el promedio de ellas como el valor buscado. Se deja al lector el ejercicio de mostrar que este valor cumple con las 2 condiciones pedidas.

Del mismo modo, tomando $q = 50$, encontramos que el segundo cuartil, o mediana, corresponde a la observación 17, y tomando $q = 75$, encontramos que el tercer cuartil corresponde a la observación 25.

Si recordamos la noción de cuartiles que habíamos introducido, nos gustaría que los valores encontrados dividieran a las observaciones en 4 grupos de, aproximadamente, igual tamaño. La Tabla IV.4 muestra lo ocurrido con las observaciones en las posiciones 9, 17 y 25, que encontramos en el conjunto de datos ficticios de 33 observaciones.

Grupo	Observaciones en el grupo, al ordenarlas de manera creciente	Número de observaciones en el grupo
1	1 a 8	8
2	10 a 16	7
3	18 a 24	7
4	26 a 33	8

Tabla IV.4: 4 grupos formados por los cuartiles, en un grupo ficticio de 33 observaciones. Las observaciones 9, 17 y 25 corresponden al primer cuartil, mediana y tercer cuartil, respectivamente, y son las que separan un grupo de otro.

La tercera columna de la Tabla IV.4 muestra que no todos los grupos contienen exactamente el mismo número de observaciones, situación que habíamos anticipado. Sin embargo, estas cantidades son bastante similares. De este modo, lo que podemos afirmar sobre los cuartiles es que ellos dividen al conjunto de observaciones, al ordenarlas de menor a mayor, en 4 grupos que contienen, aproximadamente, el mismo número de ellas, o que ellos dividen a las observaciones, al ordenarlas de menor a mayor, en 4 grupos que contienen, aproximadamente, el 25% de las observaciones. A mayor tamaño del conjunto de observaciones, mejores serán estas aproximaciones.

Si pocas observaciones del conjunto de datos, en relación al total de observaciones de este, comparten el valor de un cuartil, podemos afirmar cosas como "aproximadamente, un 25% de las observaciones son menores o iguales que el primer cuartil", o "aproximadamente un, 25% de las observaciones son mayores o iguales que el tercer cuartil", entre otras.

Para pensar

Supongamos que en un conjunto de datos encontramos que 2 cuartiles son iguales. ¿Cómo interpretaría esta situación?

Enfatizamos que el análisis que hemos realizado para encontrar los cuartiles en el ejemplo de un conjunto de datos de tamaño 33 se presenta con fines ilustrativos, y que lo más importante corresponde a la interpretación que hacemos de los valores encontrados, más que a la manera de obtenerlos.

En resumen

- Los *cuartiles* corresponden a 3 valores que dividen al conjunto de observaciones, ordenado de menor a mayor, en 4 grupos que contienen, aproximadamente, el 25% de las observaciones cada uno.
- Cuando pocas observaciones del conjunto de datos toman los valores de los cuartiles, podemos decir que, aproximadamente, un 25% de las observaciones son menores o iguales al *primer cuartil* y un 25% de las observaciones son mayores o iguales al *tercer cuartil*.
- Bajo la convención que hemos adoptado, el *segundo cuartil* corresponde a la mediana.

3.3 Otras medidas de posición relativa: quintiles, deciles y percentiles

Otras medidas de posición relativa que encontramos frecuentemente en la vida diaria corresponden a quintiles, deciles y percentiles. Estas medidas se interpretan de manera análoga a los cuartiles.

En particular, previo ordenamiento de los datos de manera creciente:

- Los *quintiles* dividen al conjunto de datos en 5 grupos con aproximadamente el mismo número de observaciones, de modo que cada uno de estos grupos contiene aproximadamente un $\frac{100}{5}\% = 20\%$ de ellas. De este modo, podemos decir, por ejemplo, que aproximadamente un 20% de las observaciones son menores o iguales al primer quintil.
- Los *deciles* dividen al conjunto de datos en 10 grupos con aproximadamente el mismo número de observaciones, de modo que cada uno de estos grupos contiene, aproximadamente, un $\frac{100}{10}\% = 10\%$ de ellas. De este modo, podemos decir, por ejemplo, que aproximadamente un 30% de las observaciones son menores o iguales al tercer decil.

- Los *percentiles* dividen al conjunto de datos en **100** grupos con, aproximadamente, el mismo número de observaciones, de modo que cada uno de estos grupos contiene aproximadamente un $\frac{100}{100}\% = 1\%$ de ellas. De este modo, podemos decir, por ejemplo, que aproximadamente un **35%** de las observaciones son menores o iguales al percentil **35**.

3.4 Medidas de posición relativa como valores puntuales, versus intervalos de valores

Discutiremos aquí diferentes ejemplos referidos a las medidas de posición relativa que hemos presentado, que enfatizan su interpretación. Haremos también el paralelo entre afirmaciones referidas a ellas, a modo de valores puntuales y a modo de intervalos.

El primer ejemplo enfatiza la interpretación de los cuartiles. Los ejemplos del **2** al **4** discuten el uso que suele darse a las medidas de posición relativa, refiriéndose a ellas como un intervalo de valores, y no valores puntuales, como aquí lo hemos presentado. Veremos que, en ambos casos, la información que se transmite es la misma.

Ejemplo 1:

En el año 2007, se reportó que el primer cuartil de ingreso familiar per cápita en Chile correspondía a alrededor de **$107.000**, la mediana a alrededor de **$182.000**, y el tercer cuartil a alrededor de **$340.000**. Esto significa que, aproximadamente:

- Un **25%** de la población percibía un ingreso familiar per cápita igual o inferior a **$107.000**.
- Un **25%** de la población percibía un ingreso familiar per cápita entre **$107.000** y **$182.000**.
- Un **25%** de la población percibía un ingreso familiar per cápita entre **$182.000** y **$340.000**.
- Un **25%** de la población percibía un ingreso familiar per cápita igual o superior a **$340.000**.

Notamos, nuevamente, que las diferencias entre los cuartiles no son las mismas. En efecto, la distancia entre el primer cuartil, **$107.000** y la mediana, **$182.000**, es de **$75.000**. Por otra parte, la distancia entre la mediana, **$182.000** y el tercer cuartil, **$340.000**, es de **$158.000**, aproximadamente el doble que la distancia anterior. Sin embargo, los porcentajes de observaciones entre los cuartiles son siempre los mismos de, aproximadamente, un **25%**.

Ejemplo 2:

Este ejemplo es una continuación del Ejemplo 1. En dicho caso, en particular, se afirma que:

> "En el año 2007, el primer cuartil de ingreso familiar per cápita en Chile correspondía a alrededor de **$107.000** y la mediana, o segundo cuartil, a alrededor de **$182.000**".

Es frecuente escuchar la misma información expresada como:

> "En el año 2007, el segundo cuartil de ingreso familiar per cápita en Chile estaba entre **$107.000** y **$182.000**".

Notamos que, en la primera afirmación, la palabra *cuartil* ha sido utilizada para referirse a valores puntuales de ingresos, **$107.000** y **$182.000**, mientras que en la segunda afirmación, la palabra *cuartil* ha sido utilizada para referirse a un intervalo de ingresos, entre **$107.000** y **$182.000**.

Para entender esta situación, la **Figura IV.22**, arriba, muestra el uso que se ha dado a la palabra *cuartil* en la primera afirmación, como valores puntuales. En ella, el primer y el segundo cuartiles son valores que determinan un intervalo que contiene, aproximadamente, un **25%** de las observaciones.

La **Figura IV.22**, abajo, muestra el uso que se le ha dado a la palabra cuartil en la segunda afirmación, como un intervalo de valores. En ella, el porcentaje de observaciones en el segundo cuartil corresponde a, aproximadamente, un **25%**.

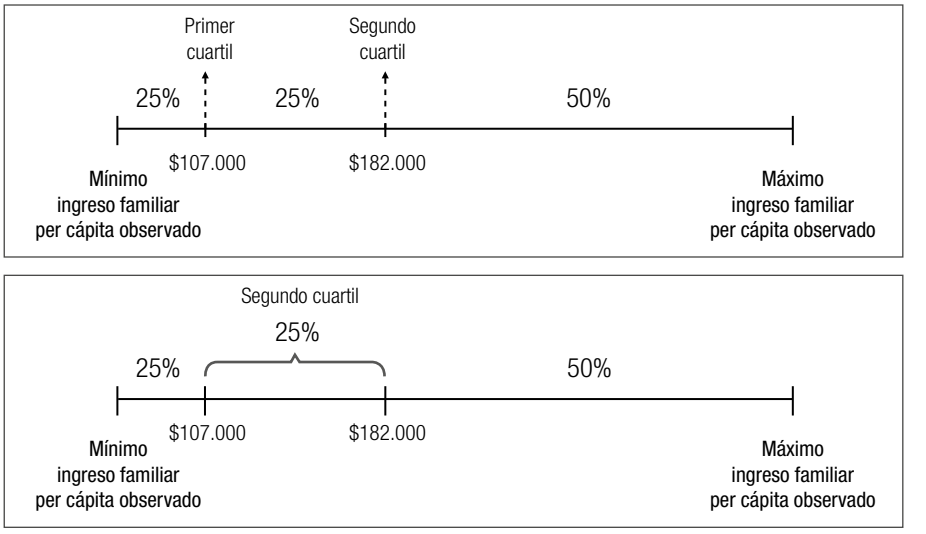


Figura IV.22: Arriba: la palabra cuartil ha sido utilizada para referirse a valores puntuales. Abajo: la palabra cuartil ha sido utilizada para referirse a un intervalo de valores.

Independientemente del uso que se le está dando al término *cuartil*, ambas afirmaciones y sus representaciones en la **Figura IV.22** comunican exactamente la misma información.

Ejemplo 3:

Consideremos la afirmación:

"El peso de Elisa se encuentra en el tercer decil de las niñas del país de su misma edad".

La **Figura IV.23** muestra la interpretación de esta afirmación, donde el término decil ha sido utilizado como un intervalo de valores. Con este mismo uso de la palabra decil, en la **Figura IV.23** observamos que:

- Como todo decil, el tercer decil contiene aproximadamente un **10%** de las observaciones.
- Existen 2 deciles con valores menores a los valores en el tercer decil. Como cada uno de estos deciles contiene aproximadamente un **10%** de las observaciones, el porcentaje total de observaciones menores o iguales a las observaciones en el tercer decil corresponde a, aproximadamente, un **2 • 10% = 20%**.
- Existen 7 deciles con valores mayores a los valores en el tercer decil. Como cada uno de estos deciles contiene aproximadamente un **10%** de las observaciones, el porcentaje total de observaciones mayores o iguales a las observaciones en el tercer decil corresponde a, apoximadamente, un **70%**.

Luego, de la afirmación realizada, podemos inferir que:

"El peso de Elisa se encuentra entre el **20%** y el **30%** de las niñas de menor peso del país y de su edad".

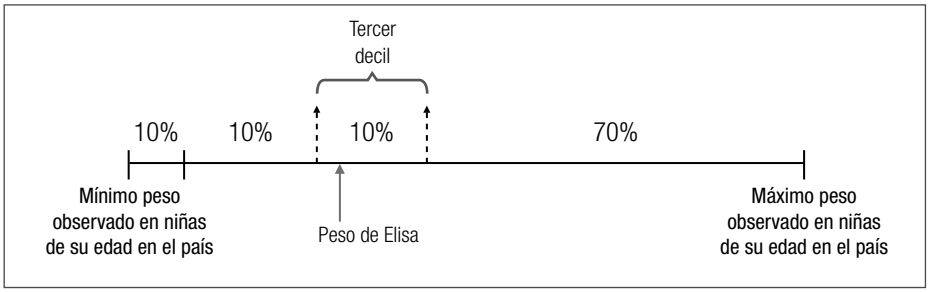


Figura IV.23: La palabra decil se ha utilizado para referirse a un intervalo de valores.

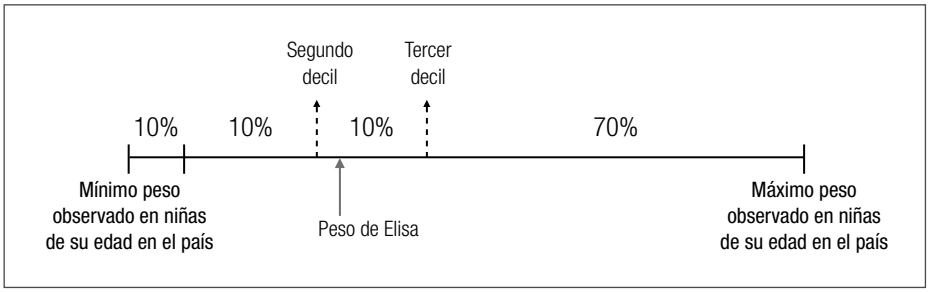


Figura IV.24: La palabra decil se ha utilizado para referirse a valores de peso puntuales.

Ejemplo 4:

Consideremos la afirmación:

"El puntaje de la escuela de Carlos en la prueba SIMCE de Lenguaje de los cuartos básicos del año 2012 se encuentra en el percentil **65** de los cursos del mismo nivel del país que rindieron dicha prueba".

Veremos paso a paso la interpretación de esta afirmación, donde el término percentil ha sido utilizado como un intervalo de valores.

- Como todo percentil, el percentil **65** contiene, aproximadamente, un **1%** de las observaciones, como muestra la **Figura IV.25.**

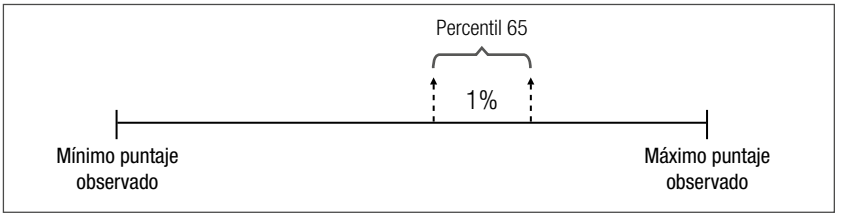


Figura IV.25: Todo percentil contiene, aproximadamente, un 1% de las observaciones del conjunto.

- Existen **64** percentiles con observaciones menores o iguales a las observaciones en el percentil **65.** Como cada uno de estos percentiles contiene, aproximadamente, un **1%** de las observaciones, el porcentaje total de observaciones menores o iguales a las observaciones en el percentil **65** corresponde a, aproximadamente, un **64 • 1% = 64%**, como muestra la **Figura IV.26.**

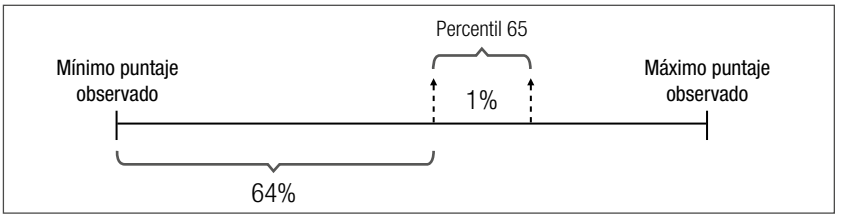


Figura IV.26: Un 64% de las observaciones es menor o igual a las observaciones en el percentil 65.

- Existen **35** percentiles con observaciones mayores o iguales a las observaciones en el percentil **65.** Como cada uno de estos percentiles contiene aproximadamente un **1%** de las observaciones, el porcentaje total de observaciones mayores o iguales a las observaciones en el percentil **65** corresponde a, aproximadamente, un **35 • 1% = 35%**, como muestra la **Figura IV.27.**

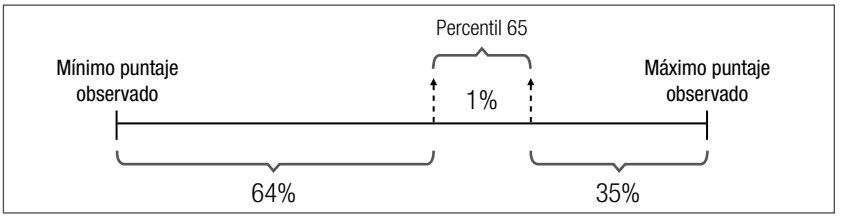


Figura IV.27: Un 64% de las observaciones es menor o igual a las observaciones en el percentil 65, y un 35% es mayor o igual a las mismas observaciones.

Luego, de la afirmación realizada podemos inferir cosas como:

"El puntaje de la escuela de Carlos en la prueba SIMCE de Lenguaje del año 2012 se encuentra dentro del **36%** superior de los puntajes obtenidos por todos los cursos del mismo nivel del país que rindieron dicha prueba".

Esto se ilustra la **Figura IV.28**.

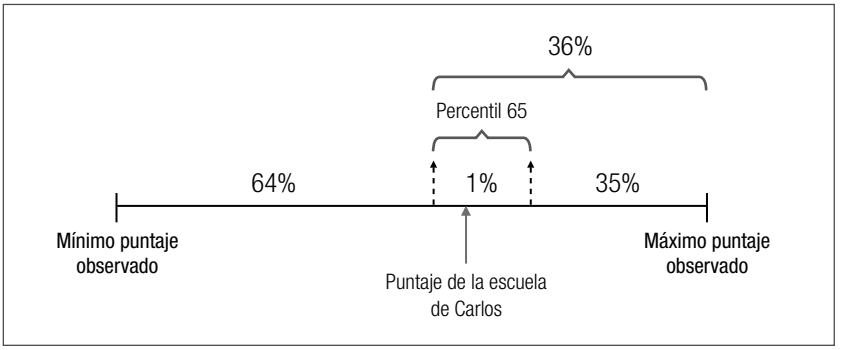


Figura IV.28: Representación de la afirmación "El puntaje de la escuela de Carlos en la prueba SIMCE de Lenguaje del año 2012 se encuentra dentro del 36% superior de los puntajes obtenidos por todos los cursos del mismo nivel del país que rindieron dicha prueba".

En resumen

- Los *cuartiles* corresponden a valores que dividen al conjunto de observaciones en **4** grupos que contienen, aproximadamente, el **25%** de las observaciones cada uno, ordenándolas de menor a mayor.
- Los *quintiles* corresponden a valores que dividen al conjunto de datos en **5** grupos que contienen, aproximadamente, el **20%** de las observaciones cada uno, ordenándolas de menor a mayor.
- Los *deciles* corresponden a valores que dividen al conjunto de datos en **10** grupos que contienen, aproximadamente, el **10%** de las observaciones cada uno, ordenándolas de menor a mayor.
- Los *percentiles* corresponden a valores que dividen al conjunto de datos en **100** grupos. El percentil p, donde p es un valor entre **0** y **100**, corresponde a un valor tal que, aproximadamente, un p% del conjunto es menor o igual a él, y un $(100 - p)$% del conjunto es mayor o igual a él.
- Los términos cuartiles, quintiles, deciles y percentiles son frecuentemente utilizados para denotar *intervalos de valores* a diferencia de *valores puntuales*.

Ejercicios

1. Utilizando observaciones recolectadas en todas sus sucursales, un banco determinó que el primer cuartil de tiempos de espera de sus clientes en la fila es de 3 minutos, y el tercer cuartil es de 8 minutos. Para cada una de las siguientes aseveraciones, indique si son siempre verdaderas. En caso de no serlo, justifique.

   a. Aproximadamente, la mitad de las personas debe esperar entre 3 y 8 min.
   b. La mediana es $\frac{(3+8)}{2} = 5{,}5$ min.
   c. Aproximadamente, el 25% de las personas debe esperar, al menos, 8 min.
   d. No es posible obtener los quintiles a partir de esta información.

2. ¿Es posible que una distribución tenga una media más alta que su tercer cuartil? Para cada una de las siguientes respuestas a esta pregunta, indique si es correcta o incorrecta. En caso de ser incorrecta, justifique.

   a. Sí, siempre es cierto.
   b. Sí, pero solo si hay valores extremos muy altos.
   c. No.

3. Los siguientes datos muestran el número de partidos de tenis que ganó un jugador cada año, durante 12 años.

   27 18 5 14 2 25 8 19 2 8 11 26

   Obtenga:

   a. Mediana.
   b. Primer y tercer cuartiles.
   c. Los números de partidos que limitan el tercer quintil (ayuda: corresponden a los percentiles 40 y 60).
   d. Interprete cada uno de los valores obtenidos en los apartados anteriores.

4. La siguiente tabla muestra las edades, en años, de un grupo de madres al nacer su primer hijo o hija.

Edad (años)	Número de madres en estudio
21	2
22	1
24	4
26	3
27	1
28	2
30	2

   a. Encuentre la mediana. Interprete este valor.

b. Encuentre los valores entre los que se encuentra el 50% central de los datos.

c. Encuentre el percentil 80. Interprételo.

5. Para cada una de las siguientes aseveraciones, indique si es verdadera o falsa. En caso de ser falsa, indique por qué.

   a. El percentil 25 corresponde a un valor tal que aproximadamente un 25% de los valores del conjunto de datos son menores o iguales a él.

   b. El segundo cuartil corresponde a la media.

   c. El percentil 75 también se conoce como tercer cuartil y se anota como Q_3.

   d. Los cuartiles Q_1, Q_2 y Q_3 dividen al conjunto de datos en 3 grupos de aproximadamente igual tamaño.

   e. El segundo quintil es un valor tal que, aproximadamente, un 40% de los datos es menor o igual a él.

   f. También se conoce como segundo quintil a un grupo consistente en un 20% del conjunto de datos, que contiene los valores entre los percentiles 20 y 40.

   g. Un cuartil siempre corresponde a un valor en el conjunto de datos.

6. Considere el siguiente diagrama de tallo y hojas, donde los valores se encuentran expresados en años, que representa las edades de las personas en una fiesta de Año Nuevo. Encuentre el valor del tercer cuartil.

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

3.5 *Boxplot*, diagrama de caja o cajón con bigotes

Hemos visto que un histograma corresponde a una representación gráfica de la distribución de las observaciones. Un *boxplot*, *diagrama de caja* o *cajón con bigotes* corresponde a una representación gráfica complementaria a un histograma, basada en la información entregada por la mediana, cuartiles y valores mínimo y máximo de un conjunto de datos. Estos 5 valores que dan origen a un *boxplot* suelen denominarse las *5 medidas de resumen* de un conjunto de observaciones.

Si bien este tipo de representaciones no está presente en el currículo escolar de Educación Básica, su comprensión contribuye a tener una visión más completa de los contenidos que un profesor debe cubrir en el aula. Por otra parte, este tipo de representaciones, así como la estadística en general, cumple un rol importante en las labores docentes y en el desempeño profesional. En efecto, a modo de ejemplo, un profesor reporta habitualmente rendimientos u otras características de sus alumnos, o recibe informes o reportes que entregan información estadística útil para su labor como docente.

3.5.1 Construcción e interpretación de un *boxplot*

Como discutimos cuando presentamos las medidas de posición relativa en general, si bien los cuartiles dividen a las observaciones en grupos de, aproximadamente, igual tamaño, esto no significa que las distancias entre ellos sean las mismas. La importancia de estas distancias es que ellas caracterizan a la distribución de las observaciones. Un *boxplot* permite visualizar esta información, como mostraremos en lo que sigue.

A modo de ejemplo, nos referiremos a las notas en una primera evaluación de los alumnos de un curso, el sexto A, cuyas 5 medidas de resumen se muestran en la **Tabla IV.5**.

Estadístico	Valor
Mínimo	3,7
Primer cuartil	4,6
Mediana	5,1
Tercer cuartil	5,4
Máximo	6,1

Tabla IV.5: 5 medidas de resumen de las notas del sexto A.

Estos valores, que recogen información respecto de la distribución de las observaciones, permiten representar a esta última en la forma que se muestra en la **Figura IV.29**, que es lo que se denomina un *boxplot*, diagrama de caja o cajón con bigotes. En la figura, el eje de las ordenadas (vertical) indica los valores de las notas. El cajón central puede tener un ancho arbitrario, pero su extensión vertical está delimitada por los cuartiles 1 y 3.

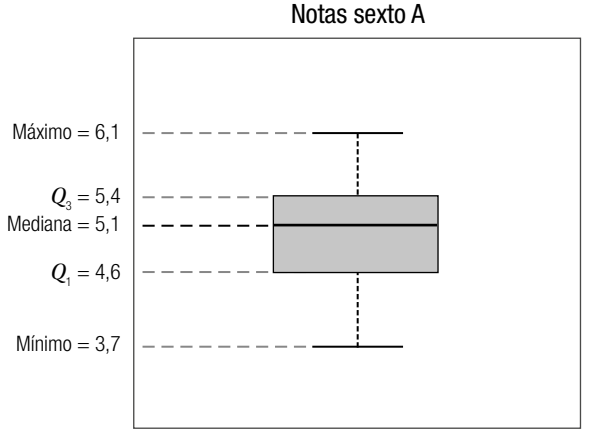


Figura IV.29: Boxplot, diagrama de caja o cajón con bigotes, de las notas del sexto A en la primera evaluación.

Dentro del mismo cajón, debe indicarse la mediana a través de una línea horizontal. Se dibuja una línea vertical, llamada "bigote", entre el primer cuartil y el valor mínimo de las observaciones, y lo mismo se hace entre el tercer cuartil y el valor máximo.

Veamos la información sobre el conjunto de datos que es posible extraer desde un *boxplot*. Como primer punto, vemos que podemos obtener información cuantitativa, como los 5 estadísticos de resumen, mínimo, máximo, primer y tercer cuartiles, y mediana.

Por otra parte, un *boxplot* también entrega información cualitativa, que es posible extraer sin la necesidad de mirar los valores en el eje de las ordenadas. En el ejemplo, en la **Figura IV.29**:

- La mediana es ligeramente más cercana al tercer cuartil que al primero. Esto dice que los valores de las observaciones se encuentran ligeramente más concentradas en la zona superior del cajón.
- Las amplitudes de los "bigotes" superior e inferior son similares. Esto nos habla de cierta similitud entre los extremos de la distribución.
- La altura del cajón es ligeramente menor a la suma de las longitudes de los "bigotes". Esto dice que los datos están levemente más concentrados en el centro.

Consideremos, como segundo ejemplo, las notas obtenidas por el sexto A en una segunda evaluación, cuyo *boxplot* se muestra en la **Figura IV.30**.

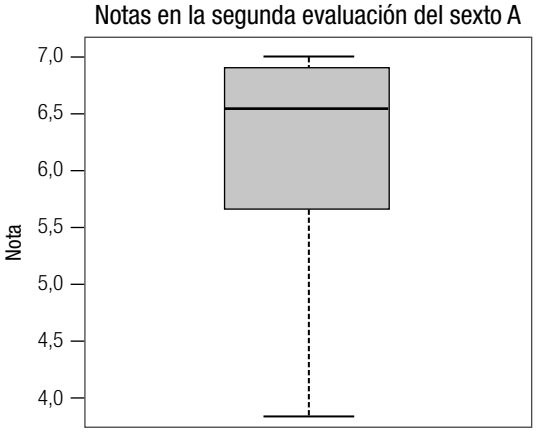


Figura IV.30: Boxplot de las notas obtenidas por el sexto A en una segunda evaluación.

Al igual que con el *boxplot* anterior, podemos obtener información cuantitativa, como los 5 estadísticos de resumen. Leemos que el mínimo fue ligeramente menor a **4**, mientras que el máximo fue aproximadamente un **7**. Por otra parte, la mediana estuvo alrededor de un **6,5**, y los primer y tercer cuartiles son, aproximadamente, **5,7** y **6,8**, respectivamente.

Alguna información cualitativa que podemos extraer corresponde a:

- La mediana no está en el centro del cajón. Esto indica que la distribución de las observaciones no es simétrica.
- La mediana está bastante más cercana al tercer cuartil que al primero. Esto dice que los valores de las observaciones se encuentran más concentradas en la zona superior del cajón, es decir, en valores altos de las notas.
- Las amplitudes de los "bigotes" superior e inferior son bastante diferentes. Esto indica que los extremos de la distribución son distintos: hay notas bajas que se alejan del grueso de los datos.
- La distribución tiene, marcadamente, una mayor concentración en las notas altas que en las bajas, lo que se observa en que la distancia entre la mediana y la nota máxima es bastante menor que la distancia entre la misma y la nota mínima.

Como indicamos anteriormente, el histograma y el *boxplot* corresponden a representaciones gráficas complementarias de la distribución de las observaciones de un conjunto. De hecho, algunos programas computacionales ofrecen la alternativa de obtener ambos gráficos de manera simultánea, como se muestra en la **Figura IV.31**, para las notas del sexto A en su segunda evaluación [3].

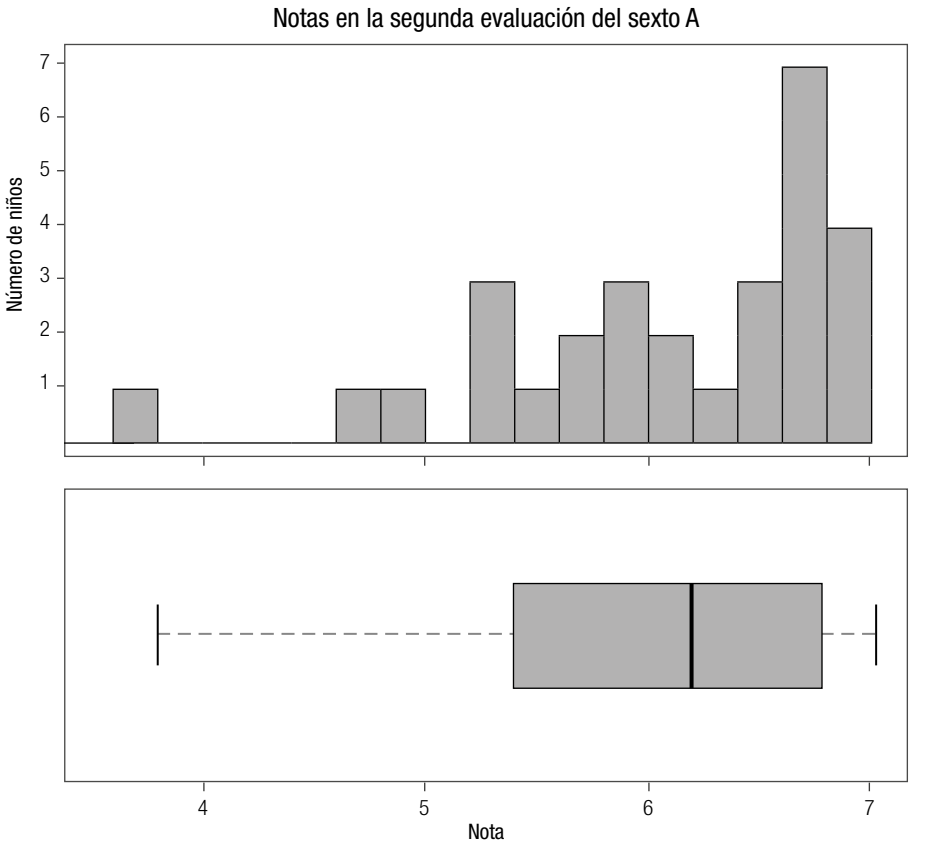


Figura IV.31: Histograma y boxplot de las notas del sexto A en su segunda evaluación.
La figura muestra que, por una parte, el histograma entrega información más detallada sobre la forma de la distribución, y, por otra, que el boxplot entrega una visión más general, indicando, por ejemplo, los valores que delimitan el 50% central de los datos, es decir, los cuartiles 1 y 3, o que la distribución posee observaciones alejadas hacia la izquierda.

3.5.2 Tratamiento de observaciones extremas en un *boxplot*

La representación de las observaciones a través de un *boxplot* también considera la presencia de observaciones extremas. Recordemos que hemos utilizado el término observación extrema para referirnos a observaciones que se alejan del grueso de los datos. Uno de los procedimientos habituales para identificar observaciones extremas se basa en relacionar la amplitud o altura del cajón, con lo que debiese ocurrir en los extremos de la distribución. Omitimos aquí los detalles de este procedimiento, y supondremos que estas observaciones ya han sido identificadas.

[3] En casos como la Figura IV.31, puede ocurrir que los extremos de los "bigotes" no coincidan con los extremos del histograma. Esto se debe a que las barras de un histograma indican intervalos de valores en los que se encuentran las observaciones, pero ellas no necesariamente se encuentran en los extremos.

En caso de existir observaciones extremas, estas se indican en el *boxplot* modificando levemente la construcción que seguimos. Para ilustrar, consideremos las notas obtenidas por el curso paralelo, el sexto B, en la primera evaluación. Las 5 medidas de resumen se muestran en la **Tabla IV.6**.

Estadístico	Valor
Mínimo	1,5
Primer cuartil	4,4
Mediana	4,8
Tercer cuartil	5,2
Máximo	5,9

Tabla IV.6 : Cinco medidas de resumen de las notas del sexto B en la primera evaluación.

Se puede mostrar que la nota **1,5** del sexto B corresponde a una observación extrema y por lo tanto, no se reporta como el extremo del "bigote" inferior. Esta observación se identifica en el *boxplot* a través de un punto, como se muestra en la **Figura IV.32**, acortándose además la extensión del "bigote" inferior. En presencia de observaciones extremas en el extremo superior de la distribución, estas también deben indicarse a través de puntos, acortándose, en ese caso, el "bigote" superior.

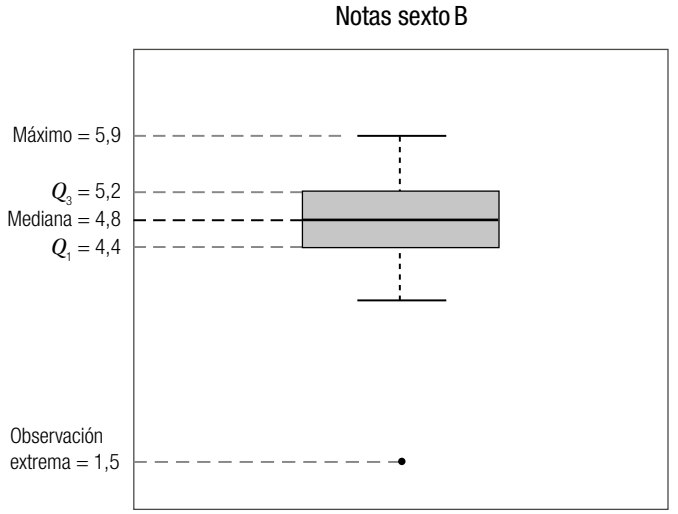


Figura IV.32: Boxplot de las notas obtenidas por el sexto B. Se muestra una observación extrema.

La idea de aislar las observaciones extremas en la figura es que podamos describir el comportamiento general de los datos, y esto corresponde a describir y analizar lo que ocurre con el grueso de ellos.

En la **Figura IV.32** notamos que:

- El "bigote" inferior no llega hasta la observación mínima, debido a que esta es considerada una observación extrema. En su lugar, este "bigote" se extiende hasta la menor observación, cuyo valor es mayor o igual a una cota establecida.
- La observación identificada como observación extrema se muestra a través de un punto en la parte baja de la figura, a la altura de su valor, **1,5**.

- Haciendo un paralelo entre el *boxplot* y el histograma de este mismo conjunto de datos, como se muestra en la **Figura IV.33**, notamos que el histograma posee una observación bastante pequeña, que se identifica como observación extrema en el *boxplot*.

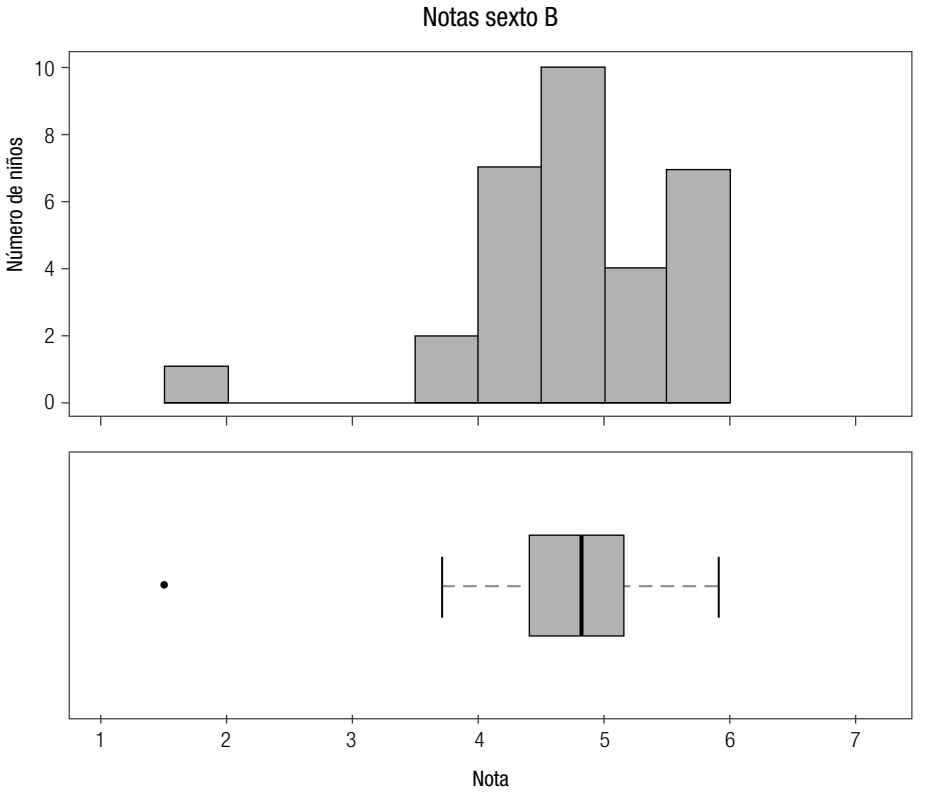


Figura IV.33: Histograma y boxplot de las notas del sexto B. Se identifica una observación extrema inferior.

En resumen

- Un *boxplot*, *diagrama de caja* o *cajón con bigotes* corresponde a una representación gráfica de la distribución de las observaciones.
- En un *boxplot* representado verticalmente, se observa una caja cuyo límite superior corresponde al *tercer cuartil* y el inferior al *primer cuartil*. La *mediana* es identificada dentro del cajón a través de una línea horizontal.
- En general, los "bigotes" se extienden hasta los valores *mínimo* y *máximo*.
- Una mayor amplitud vertical de la caja implica mayor variabilidad.
- Una mayor longitud de los "bigotes" indica mayor variabilidad.
- En caso de existir observaciones indicadas a través de puntos más allá de los extremos de los "bigotes", corresponden a observaciones que pueden ser consideradas *observaciones extremas*.

3.5.3 Uso de *boxplots* para comparar distribuciones

Supongamos que nos interesa comparar los rendimientos de los cursos sexto A, B y C, en la primera evaluación. Podemos hacerlo representando las notas de los 3 cursos en *boxplots* paralelos, como se muestra en la **Figura IV.34**.

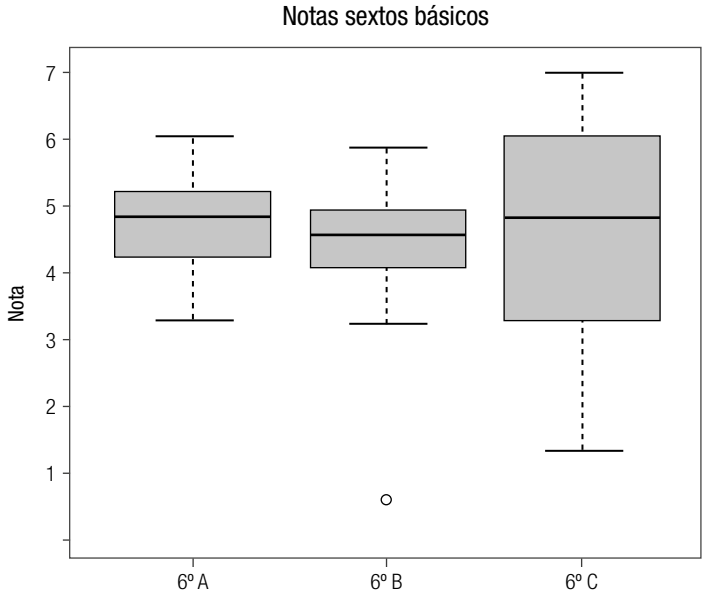


Figura IV.34: Boxplots de las notas de los sextos A, B y C.

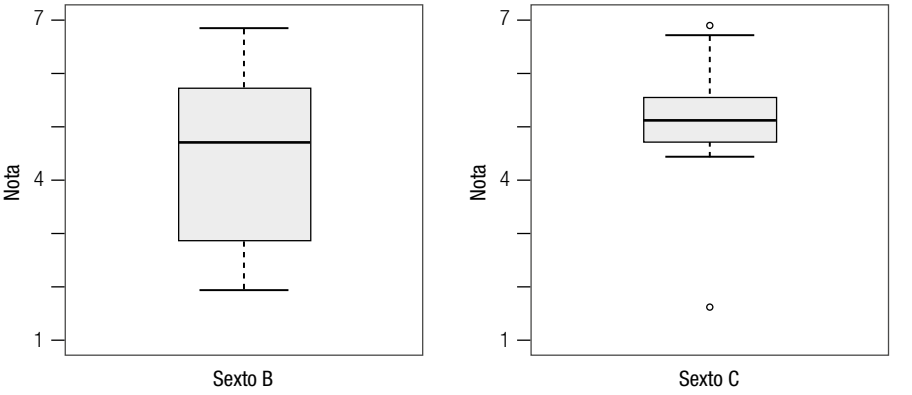
Una primera observación que se puede hacer a partir de la **Figura IV.34** es que existe una observación extrema en el sexto B. Esto significa que existe un niño que obtuvo una nota muy por debajo de las de sus compañeros. Por otra parte, en la figura observamos que la mediana de las notas del sexto B es ligeramente menor a las medianas de los sextos A y C, que corresponden a, aproximadamente, un **5,0**. Por otra parte, tanto el **50%** central de las notas del sexto A, como de las del sexto B, se mueven en intervalos de valores de menos de 1 punto (**10** décimas) de ancho: entre aproximadamente **4,6** y **5,2** para el sexto A, y **4,5** y **5,1** para el sexto B. Sin embargo, el **50%** central de las notas del sexto C se mueve en un intervalo de alrededor de **2** puntos (**20** décimas) de ancho, mostrando que las notas de sus alumnos se encuentran más alejadas de la mediana que las notas de los alumnos de los cursos A y B. También notamos que en el sexto C la nota más baja, o el mínimo, es bastante menor a las notas más bajas de los sextos A y B (excluyendo la observación extrema en el sexto B). Sin embargo el sexto C también posee al menos una nota cercana a **7,0**, lo que no ocurre con los sextos A y B.

Podemos también notar que las distribuciones de las notas de los sextos A y B son bastante simétricas en torno a la mediana, no así la distribución de las notas del sexto C, que posee una ligera cola hacia los valores pequeños. Esta apreciación se basa en que el "bigote" inferior de las notas del sexto C es de mayor longitud que el "bigote" superior.

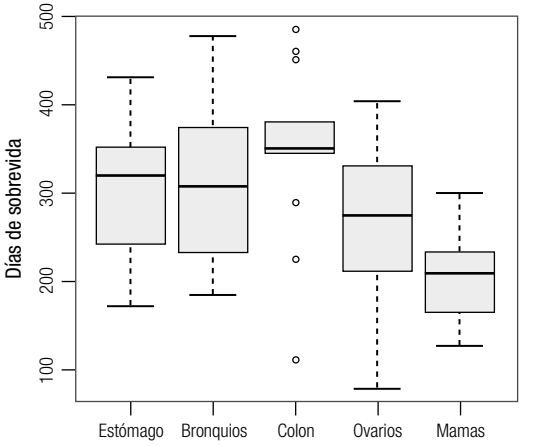
Es importante notar que, para realizar las comparaciones de manera adecuada, se debe cuidar que todos los *boxplots* utilicen la misma escala en el eje vertical o de las ordenadas. De este modo, es posible comparar visualmente tanto los valores de los **5** estadísticos de resumen, como también comparar apreciaciones sobre las formas de las distribuciones.

Ejercicios

1. Realice una comparación de las distribuciones de las notas finales obtenidas por los sextos B y C, en base a los siguientes *boxplots*.



2. En cierto estudio, interesa determinar si las distribuciones de los tiempos de sobrevida al cáncer de pacientes bajo un mismo tratamiento son diferentes dependiendo del órgano que esté involucrado. Para esto, se tomó una muestra de pacientes de oncología de cierto centro médico y se obtuvieron los gráficos de caja que se muestran a continuación. El eje de las ordenadas (vertical) indica el número de días de sobrevida.



   a. Obtenga de manera aproximada las medianas de los tiempos de sobrevida, por tipo de cáncer. Compárelas y comente.

   b. Observe y compare las amplitudes de los rangos de valores que corresponden al 50% central de los datos, para cada tipo de cáncer. ¿Qué puede decir sobre la cercanía o lejanía de estos valores con respecto a la mediana?

   c. ¿Cómo interpreta el hecho de que el *boxplot* de los tiempos de sobrevida al cáncer de colon no posea "bigotes"?

   d. Compare de manera general las distribuciones de los tiempos de sobrevida de los diferentes tipos de cáncer.

4. Medidas de dispersión

Para introducir las ideas, trabajaremos con los 3 conjuntos de notas obtenidas por los 3 sextos básicos paralelos en una misma evaluación. Los datos se presentan mediante histogramas, que se muestran en la **Figura IV.35**.

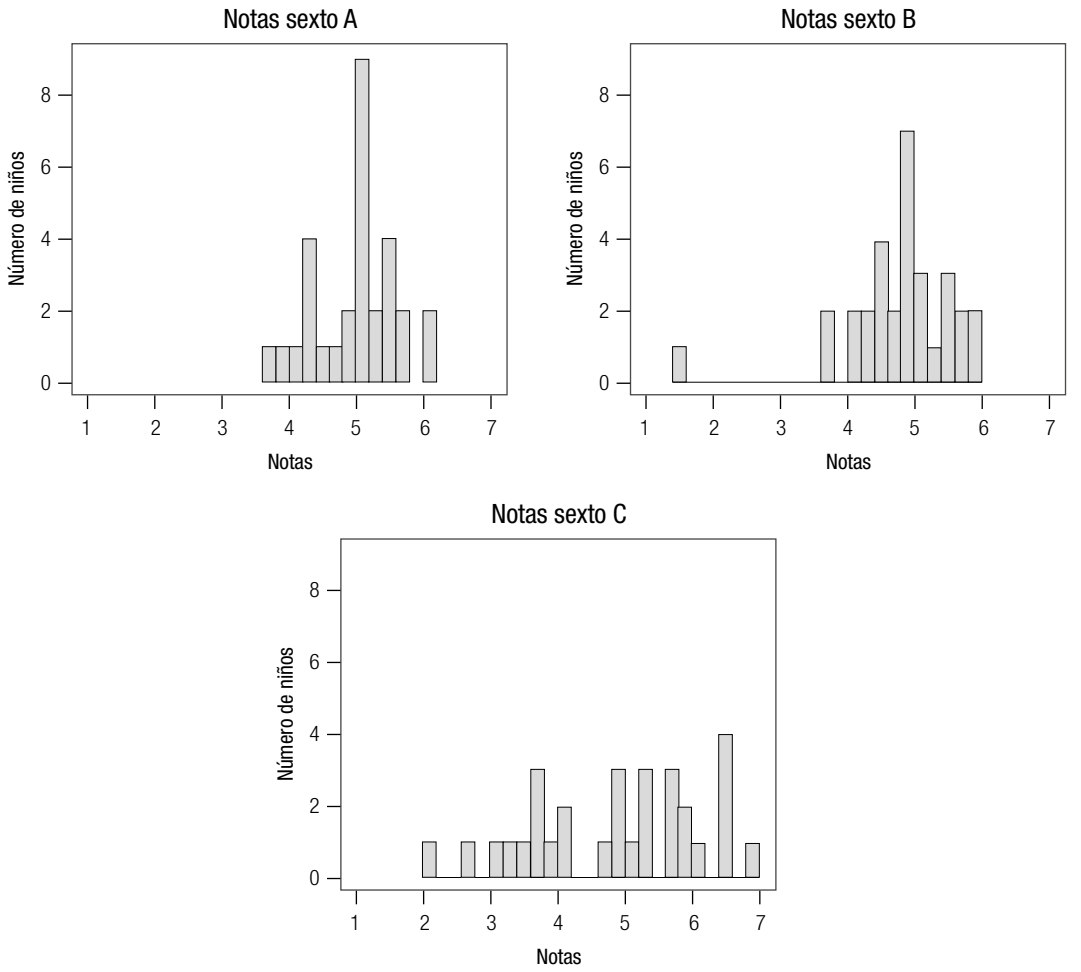


Figura IV.35: Notas obtenidas por los tres sextos básicos de una escuela, en una misma evaluación.

Para pensar

¿A qué curso le fue mejor? ¿en qué curso las notas fueron más homogéneas? ¿por qué?

Consideremos los rendimientos en los sextos A y C. Para comparar las notas de ambos cursos, se propone calcular sus medidas de tendencia central. Utilizando las notas individuales de los alumnos (no se muestran), se obtiene que la media y la mediana del sexto A corresponden a **5,0** y **5,1**, respectivamente, y que media y mediana del sexto C corresponden a **4,9** y **5,0**, respectivamente.

Según las medidas de resumen calculadas, las distribuciones de ambos cursos son similares en cuanto a su centro. Sin embargo, en la figura observamos que ambas distribuciones son bastante diferentes. En el sexto A, las notas están relativamente concentradas en torno al centro de la distribución, mientras que en el sexto C están, en general, bastante más alejadas de este. Probablemente, aunque las notas del sexto A y sexto C posean medidas de tendencia central similares, la profesora del sexto C estará más preocupada por el rendimiento de su curso que la profesora del sexto A, al notar, por ejemplo, que un mayor número de sus alumnos obtuvo notas bajo **4,0**.

Por otra parte, la **Figura IV.35** también sugiere que el rendimiento del sexto A fue muy parecido al del sexto B, dado que, en general, las notas de ambos cursos se mueven en intervalos de valores similares. Sin embargo, en el sexto B existe una nota sustantivamente menor que, probablemente, también preocupará a su profesora.

Quisiéramos disponer de estadísticos que dieran cuenta de estas características de los conjuntos de datos: su dispersión o variabilidad. Los estadísticos que cumplen esta misión se denominan *medidas* o *estadísticos de dispersión*, y estudiaremos 3 de ellos: el recorrido, el recorrido intercuartil y la desviación estándar. Hacemos notar que estas medidas de dispersión solo tienen sentido para variables cuantitativas.

4.1 Recorrido

Consideremos nuevamente las notas de los sextos A y C, en la **Figura IV.35**. Notamos que las notas del sexto A se ubican en un intervalo de valores de menor amplitud que las notas del sexto C, es decir, las notas máxima y mínima del sexto A son más cercanas entre sí que las notas máxima y mínima del sexto C. Esto nos transmite la idea que queremos capturar, de que las notas en el sexto A son menos variables, o más homogéneas, que las del sexto C.

La diferencia entre los valores máximo y mínimo de un conjunto de datos, que corresponde a la amplitud del intervalo que estos recorren, se denomina *recorrido* de las observaciones o de la distribución de estas, y corresponde a una medida de la dispersión, o variabilidad, de la distribución. Hacemos notar que el currículo escolar hace referencia a esta medida llamándola *rango*.

La **Tabla IV.7** muestra que las notas mínima y máxima del sexto A corresponden a **3,7** y **6,1**, respectivamente. De este modo, el recorrido de las notas del sexto A, que corresponde a la diferencia entre estas dos cantidades, es **6,1 – 3,7 = 2,4** puntos. Repitiendo el procedimiento para los cursos sexto B y sexto C se obtienen los resultados restantes en la **Tabla IV.7**.

Curso	Mínimo	Máximo	Recorrido
6º A	3,7	6,1	6,1 – 3,7 = 2,4
6º B	1,5	5,9	5,9 – 1,5 = 4,4
6º C	2,1	6,9	6,9 – 2,1 = 4,8

Tabla IV.7: Valores máximo, mínimo y recorrido de las notas de los cursos paralelos, sextos A, B y C.

El recorrido de las notas del sexto A, **2,4** puntos, es el menor de todos, indicando una menor dispersión de las observaciones que las de los cursos B y C, que es lo que muestran los histogramas de las notas en la **Figura IV.35**. En la misma figura, se observa que el grueso de las notas de sexto A

se encuentra entre valores muy similares a los valores entre los que se encuentra el grueso de las notas del sexto B. Sin embargo, en la **Tabla IV.7** se observa que el recorrido de las notas del segundo curso es mayor, dando cuenta de la observación extrema a la izquierda en la **Figura IV.35**. Los recorridos de sexto B y sexto C son bastante similares, **4,4** y **4,8** puntos, respectivamente.

En general, mientras mayor es el recorrido de una distribución, más dispersa o más variable es esta.

En resumen

- El *recorrido* de las observaciones corresponde a una medida de la dispersión de estas, y se refiere a la diferencia entre el máximo y el mínimo del conjunto de observaciones.
- Mientras mayor es el recorrido de un conjunto de observaciones, mayor es su dispersión o variabilidad.

4.2 Recorrido intercuartil

Dado que los valores mínimo y máximo de un conjunto de datos cambian instantáneamente si se introducen observaciones extremas, el recorrido resulta muy sensible a la presencia de estas últimas: basta una sola observación muy grande o muy pequeña para que el recorrido aumente sustancialmente. Decimos, entonces, que el recorrido es poco robusto frente a valores extremos.

Esta situación se evidenció al estudiar el comportamiento de las notas del sexto B en la **Figura IV.35**, donde, si bien en general las notas están bastante concentradas, la presencia de una nota extrema pequeña hace crecer el recorrido desde aproximadamente **2,5** puntos sin dicha nota, a **4,4** puntos al incluirla.

La estabilidad de la mediana y de los cuartiles frente a valores extremos sugiere utilizar estadísticos de dispersión basados en estas medidas, debido a que, por su construcción, serán más robustos frente a dichos valores. Una medida de variabilidad de los datos que cumple con estos requisitos corresponde a la distancia entre los cuartiles **1** y **3**, Q_1 y Q_3, estadístico que se denomina *recorrido intercuartil* o *RIC*. Vemos que el recorrido intercuartil es una medida de la dispersión del **50%** central de los datos, al ordenarlos según su magnitud:

$$\text{RIC} = Q_3 - Q_1$$

A modo de ejemplo, el primer y tercer cuartil de las notas del sexto A son **4,7** y **5,1**, respectivamente. De este modo el recorrido intercuartil del sexto A corresponde a:

$$\text{RIC} = 5{,}1 - 4{,}7 = 0{,}4$$

Es decir, **4** décimas. Por otra parte, se puede obtener que el recorrido intercuartil de las notas del sexto C corresponde a **2,6** puntos, mostrando una variabilidad considerablemente mayor.

Notemos que el recorrido intercuartil corresponde a una medida de la dispersión de las observaciones alrededor de la mediana, dado que indica la amplitud de un intervalo de valores alrededor de ella, que contiene al 50% central de los datos. Una mayor amplitud o un mayor rango intercuartil indica que nos debemos alejar más de la mediana para abarcar el 50% de las observaciones. Esto indica una mayor variabilidad.

En resumen

- El *recorrido intercuartil* corresponde a una medida de variabilidad o dispersión de las observaciones, y se refiere a la distancia entre los cuartiles 1 y 3.
- Mientras mayor es el recorrido intercuartil, mayor es la variabilidad o dispersión de las observaciones.

4.3 Desviación típica o estándar [4]

Así como el recorrido intercuartil mide dispersión de las observaciones con respecto a la mediana, la *desviación estándar* corresponde a un estadístico que mide la dispersión con respecto a la media. Para comprenderlo, seguiremos el razonamiento que lleva a su construcción:

1. Para cada observación, su distancia a la media corresponde a una *medida de su dispersión individual* con respecto a esta. Una observación alejada de la media se considera más dispersa que una observación cercana.

   A modo de ejemplo, en las notas del sexto A, la media corresponde a 5,0. Una nota de 4,2 se encuentra a 8 décimas de la media, mientras que una nota 4,9 se encuentra a 1 décima de la misma. Consideramos que la primera nota es más dispersa con respecto a la media que la segunda.

   La segunda columna de la Tabla IV.8 muestra las dispersiones individuales de cada una de las notas del sexto A, obtenidas como la diferencia entre cada una de ellas y la media del curso, 5,0.

Nota	dispersión individual (en puntos)	$(\text{dispersión})^2$ (en puntos al cuadrado)
3,7	$(3{,}7 - 5{,}0) = -1{,}3$	$(-1{,}3)^2 = 1{,}69$
3,8	$(3{,}8 - 5{,}0) = -1{,}2$	$(-1{,}2)^2 = 1{,}44$
4,0	−1,0	1,00
…	…	…
…	…	…
6,1	1,1	1,21
6,1	1,1	1,21
Media = 5,0 (puntos)		Suma = 11,09 $(\text{puntos})^2$

Tabla IV.8: Ilustración de pasos intermedios en la obtención de la desviación estándar de las notas del sexto A.

[4] Al igual que los contenidos asociados a *boxplots*, la desviación típica o estándar no pertenece al currículo escolar de primero a sexto básico.

2. Quisiéramos que dos observaciones ubicadas a la misma distancia de la media, aunque una sea menor que ella y la otra mayor, tuvieran la misma medida de dispersión individual. A modo de ejemplo, dos notas, 4,8 y 5,2, se encuentran ambas a 2 décimas de la media, sin embargo, sus dispersiones son –0,2 y 0,2.

   Para superar este inconveniente, estas cantidades son elevadas al cuadrado, con lo que se pierde el concepto de ser menor o mayor que la media (puesto que $(-0{,}2)^2 = (0{,}2)^2 = 0{,}04$), pero se mantiene la idea de que mayores magnitudes indican mayores dispersiones.

   La tercera columna de la **Tabla IV.8** muestra estas cantidades, obtenidas al elevar al cuadrado las desviaciones individuales.

3. Para resumir el comportamiento de dispersión de todas las observaciones, podríamos tomar el promedio de las desviaciones individuales elevadas al cuadrado, sumándolas y dividiendo esta suma por el número total de observaciones.

   Es posible mostrar que existen ventajas al dividir la suma de las desviaciones al cuadrado por el número de observaciones menos una, en lugar de por el número total de ellas. La justificación incluye una discusión que escapa a los contenidos de este texto, sin embargo, tomaremos esta convención.

   En el ejemplo, el sexto A tiene 30 niños, por lo que, utilizando la tercera columna de la **Tabla IV.8**, el cálculo se haría como:

$$\frac{1{,}69 + 1{,}44 + 1 + \ldots + 0{,}64 + 1{,}21 + 1{,}21}{30 - 1} = \frac{11{,}09}{29} = 0{,}38 \text{ puntos}^2$$

   La cantidad así obtenida se denomina *varianza* de la distribución. Un mayor valor de este estadístico indica un conjunto de datos con mayor dispersión o variabilidad.

4. Notamos que la varianza está expresada en puntos al cuadrado, puesto que hemos sumado las desviaciones al cuadrado, lo que dificulta su interpretación. Una medida más simple de interpretar debiese estar expresada en las unidades originales de los datos. Debido a esto, se utiliza la raíz de la varianza como estadístico de dispersión, el que se denomina *desviación típica o estándar*. En este caso, esta corresponde a:

$$\sqrt{0{,}38\,\text{puntos}^2} = 0{,}62\,\text{puntos}$$

Si repetimos el análisis para las notas de los cursos restantes, obtenemos los valores de la **Tabla IV.9**.

Curso	Varianza $(\text{puntos})^2$	Desviación estándar (puntos)
Sexto A	0,38	0,62
Sexto B	0,71	0,84
Sexto C	1,63	1,28

Tabla IV.9: Varianzas y desviaciones estándar de las notas de los cursos paralelos, sextos A, B y C.

De la **Tabla IV.9** y los histogramas en la **Figura IV.35**, podemos destacar:

- Las distribuciones de las notas de los sextos A y B son similares, ya que la segunda tiene una observación extrema a la izquierda. Esto hace crecer su desviación estándar hasta **0,84**. De hecho, si no se tuviese dicha observación extrema, la desviación estándar del sexto B sería de **0,60**, cercana a la desviación estándar del sexto A.
- Las distribuciones de las notas de los sextos B y C tienen recorridos muy similares. Sin embargo, la segunda tiene mayor desviación estándar debido a que las notas se distribuyen uniformemente en el intervalo de valores, mientras que en la primera, ellos se encuentran bastante más concentrados.

En resumen

- La *desviación estándar* de un conjunto de datos corresponde a una medida de la dispersión de los datos con respecto a la media, es decir, de qué tan alejados se encuentran los datos de esta última.
- Una mayor desviación estándar indica una mayor variabilidad en las observaciones.

4.4 Errores y dificultades relacionadas a medidas de dispersión

Así como ocurre con las medidas de tendencia central, muchos de los errores y dificultades en el aprendizaje de medidas de dispersión se deben a la actitud equivocada de enfocarse en el procedimiento para encontrar sus valores, y no en lo que ellas informan acerca de los datos. A continuación, mencionamos algunos errores y dificultades:

- *Reportar el recorrido, o el recorrido intercuartil, como un intervalo de valores.* Recordemos que estas dos medidas se utilizan para expresar la variabilidad o dispersión de las observaciones. Supogamos, a modo de ejemplo, que en un grupo de niños el peso mínimo corresponde a 45 kilos y el máximo, a 52 kilos. Al referirse al recorrido sería erróneo decir que "el recorrido de los pesos de los niños es entre 45 y 52 kilos". Esta afirmación entrega información sobre los valores de las observaciones y no sobre qué tan dispersas o alejadas están entre sí. En este ejemplo, el recorrido de las observaciones es $52 - 45 = 7$ kilos.

- *Calificar un conjunto de datos como disperso debido a que su histograma es disparejo*. La **Figura IV.36** ilustra esta idea. En ella, el histograma es irregular o disparejo en el sentido de que las alturas de barras contiguas son bastante diferentes. Sin embargo, la dispersión se refiere a los valores de las observaciones, en el eje horizontal, y no a sus frecuencias, en el eje vertical.

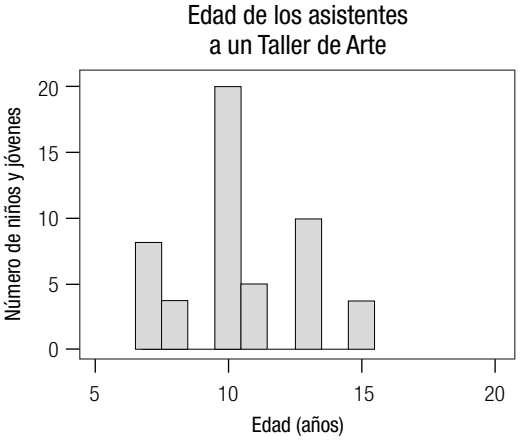
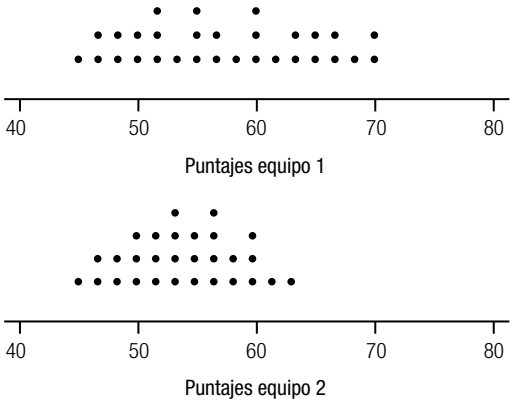


Figura IV.36: Las diferencias existentes entre las alturas de las barras no significan que las observaciones sean más o menos dispersas. La dispersión está dada por los valores de las observaciones, en el eje horizontal o de las abscisas.

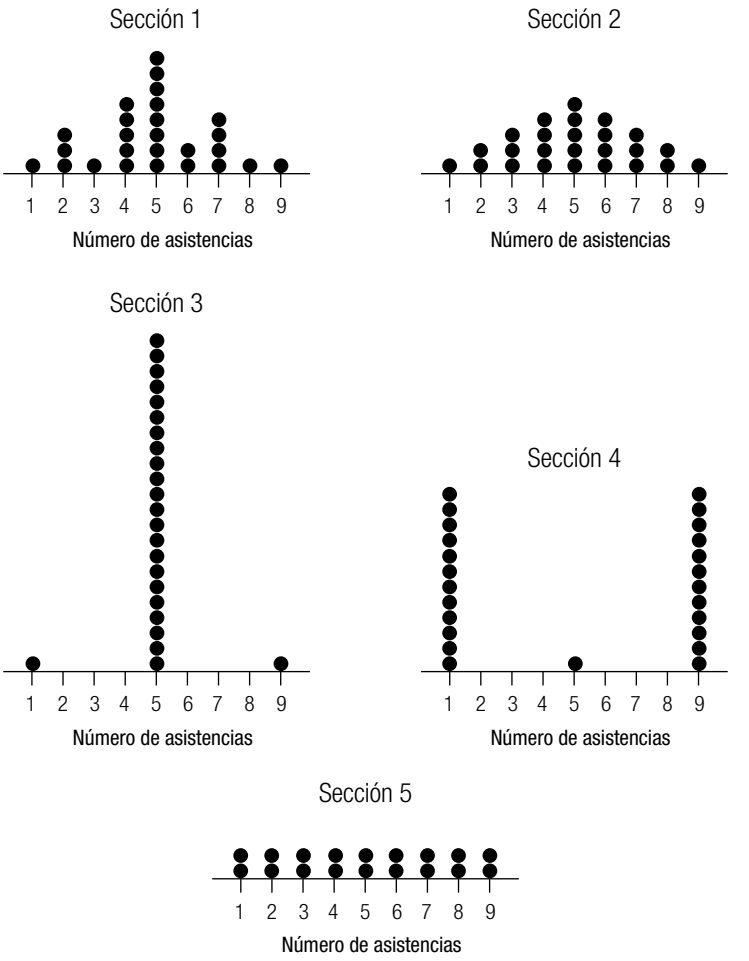
Ejercicios

1. Los siguientes gráficos muestran los puntajes, en una escala de **1** a **100**, obtenidos por dos equipos de tiro al blanco:



   a. Obtenga el recorrido de los puntajes de cada uno de los dos equipos. De acuerdo a estos, ¿qué equipo tuvo un comportamiento más homogéneo o menos variable?
   b. Sin obtener los valores, ¿qué equipo cree usted que tiene la menor desviación estándar en sus puntajes? Explique. De acuerdo a esto, ¿qué equipo tuvo un comportamiento más homogéneo o menos variable?
   c. Utilice a. y b. para comparar los comportamientos de los dos equipos.

2. En 5 secciones diferentes de un curso sobre apreciación musical, se pidió a los alumnos que indicaran el número de veces que habían asistido a conciertos de ópera. Las siguientes figuras muestran las distribuciones de los números de asistencias, por sección.



a. ¿Qué sección diría usted que tiene mayor variabilidad? Justifique.

b. ¿Qué sección diría usted que tiene la menor variabilidad? Justifique.

c. Obtenga el recorrido intercuartil y la desviación estándar de cada sección, e indíquelos en la siguiente tabla:

	Sección 1	Sección 2	Sección 3	Sección 4	Sección 5
Recorrido intercuartil					
Desviación estándar					

d. Entre las secciones 1 y 2, ¿qué diagrama de puntos es más disparejo? ¿Es la variabilidad de esa sección mayor o menor que la de la otra? Justifique.

e. Entre las secciones 3, 4 y 5, ¿qué distribución tiene el mayor número de valores diferentes? ¿Es la variabilidad de esa sección la mayor de las variabilidades de las 3 secciones? Justifique.

f. Basado en sus respuestas anteriores, describa como lo disparejo del diagrama de puntos y la variedad de los valores no están directamente relacionados al concepto de variabilidad.

g. Elabore un ejemplo hipotético del número de asistencias a conciertos de 20 nuevos integrantes del taller de apreciación musical, cada uno entre 1 y 9 asistencias, inclusive, permitiendo repetidos, de tal manera que se produzca una desviación estándar tan pequeña como sea posible.

h. Repita el ejercicio anterior elaborando datos con una desviación estándar tan grande como sea posible.

3. Las siguientes corresponden a las edades, en años, de los 8 alumnos de un curso que fueron a una salida a terreno, y la edad de su profesor.

20 19 19 25 20 18 19 52

a. Encuentre el recorrido de los datos.

b. Encuentre la media de la distribución de las edades en este conjunto de datos y utilícela para llenar la siguiente tabla. Obtenga, a partir de ella, la desviación estándar de los datos.

Asistente	Edad (años)	Dispersión individual (edad - media)	(Dispersión individual)^2
1	20		
2	19		
3			
4			
5			
6			
7			
8	52		
	Media =		Suma =

c. ¿Qué medida de dispersión, recorrido o desviación estándar, es preferible utilizar en este caso?

4. Los siguientes datos corresponden a los pesos de 10 recién nacidos, medidos en kilogramos:

2,977 3,155 3,920 3,412 4,236 2,593 3,270 3,813 4,042 3,387

a. Obtenga el recorrido de los pesos de estos recién nacidos e interprételo en palabras.

b. Encuentre la media de la distribución de los pesos en este conjunto de datos, y utilícela para llenar la siguiente tabla. Obtenga, a partir de ella, la desviación estándar de los datos.

Recién nacido	Peso al nacer (kg)	Dispersión individual (peso - media)	(Dispersión individual)^2
1	2,977		
2	3,155		
3			
4			
5			
6			
7			
8			
9			
10	3,387		
	Media =		Suma =

5. Los siguientes datos corresponden a los tiempos, en minutos, utilizados en el transporte para llegar a su lugar de trabajo de **12** trabajadores en un día en particular.

18 34 68 22 10 92 46 52 38 29 45 37

a. Obtenga el primer y el tercer cuartil de estos tiempos. Utilícelos para obtener el recorrido intercuartil de los tiempos.

b. Suponga que, al día siguiente, el trabajador cuyo tiempo de transporte había sido de 68 minutos se encontró un accidente en la ruta, por lo que demoró **95** minutos. ¿Cómo cambia el recorrido intercuartil, si los restantes trabajadores demoraron lo mismo que el primer día?

6. Los siguientes números corresponden a la cantidad de pacientes diarios que consultaron un centro médico, durante los últimos **30** días.

83	64	84	76	84	54	75	59	70	61
63	80	84	73	68	52	65	90	52	77
95	36	78	61	59	84	95	47	87	60

a. Obtenga el recorrido de la distribución de los datos.

b. Obtenga la desviación estándar de la distribución de los datos.

c. Obtenga el recorrido intercuartil de la distribución de los datos.

d. ¿Cómo cambiaría el recorrido de los datos si al día siguiente visitaran el centro **150** pacientes?

e. ¿Cómo cambiaría el recorrido intercuartil de los datos si en el día siguiente visitaran el centro 150 pacientes?

f. ¿De qué manera cree que cambiaría la desviación estándar en ese mismo caso?

g. ¿Qué puede concluir a partir de sus respuestas anteriores?

Ejercicios del capítulo

1. Las siguientes observaciones corresponden a los pesos, expresados en kilogramos, de los alumnos de un quinto básico:

54	63	40	57	48	62	46	43	45	49	40
37	41	61	32	58	62	54	58	60	58	59
52	59	43	55	45	38	63	49	51	51	59
56	37	41	50	48	57	51	49	51	55	47

   a. Determine las medidas de tendencia central de los pesos de los alumnos. Compare sus valores e interprete en términos de la forma de la distribución.

   b. Determine los cuartiles 1 y 3 de los pesos de los alumnos. Interprételos.

   c. Determine el valor del primer quintil de los pesos de los alumnos. Interprételo.

   d. Encuentre el intervalo de valores denominado segundo quintil de los pesos de los alumnos. Interprételo.

2. Considere la información de la siguiente tabla:

Deporte favorito	Alumnos
Fútbol	6
Básquetbol	4
Gimnasia rítmica	3
Tenis	2

   ¿Qué medida(s) de tendencia central es razonable determinar en este caso? Justifique su respuesta.

3. Una fábrica de cajas de cartón tiene 50 trabajadores y un gerente. Cuando se agrega el sueldo del gerente, quien tiene un pago muy superior al de los trabajadores, ¿qué medida de tendencia central de los sueldos de todos los empleados es más estable, al considerar, o no, el sueldo del gerente? Justifique su respuesta.

4. La siguiente tabla muestra las alturas de los alumnos de un curso, expresadas en metros, y agrupadas en intervalos:

Altura (m)	Número de alumnos
1,33-1,36	3
1,37-1,40	2
1,41-1,44	6
1,45-1,48	6
1,49-1,52	10
1,53-1,56	3
1,57-1,60	1

Agregue a la tabla una columna que indique las frecuencias relativas porcentuales, y responda las siguientes preguntas:

a. ¿Qué porcentaje de alumnos miden entre **1,49** y **1,52 m**?
b. ¿Cuál es el intervalo de valores más angosto que puede determinar, a partir de la tabla, en el que se encuentra la mediana de las alturas de los alumnos?
c. ¿Cuál es el intervalo de valores más angosto que puede determinar, a partir de la tabla, en el que se encuentra el percentil **90** de las alturas de los alumnos?

5. Un profesor plantea a sus alumnos el siguiente problema:

> "Un grupo de 10 niños en una fiesta de cumpleaños tiene, en promedio, **10,5** años. ¿Cuál será el promedio de edad de estos mismos niños en, exactamente, un año más?".

Uno de sus estudiantes dice que el promedio no variará. ¿Por qué cree usted que el alumno comete un error al responder? Justifique su respuesta.

6. Las siguientes corresponden a propiedades de la media de todo conjunto de observaciones:

i. Si se multiplica la media por el número total de observaciones, se obtiene la suma de estas.
ii. Si a cada una de las observaciones se le resta la media y luego se suman estas diferencias, la suma resultante es igual a **0**.
iii. Si se suma o resta una misma cantidad a cada una de las observaciones, la media se verá aumentada o disminuida en dicha constante.
iv. Si se multiplica o divide cada una de las observaciones por una constante (diferente de **0**, en el caso de la división), el promedio se verá multiplicado o dividido por dicha constante.

Muestre que estas propiedades se cumplen, en particular, en el conjunto de los números de lápices en los estuches de **5** niños:

12 7 15 9 23

En los apartados iii. y iv., utilice la constante igual a **2**.

7. La siguiente actividad simula ser tomada de un texto escolar:

Un joven practicante debe llevar a cabo un control de calidad de los tornillos fabricados por una industria metalmecánica. Para esto, dispone de la siguiente tabla con los largos de los tornillos en una muestra representativa del producto terminado:

Largo tornillos (mm)	8	9	10	11	12
Nº de tornillos	18	35	28	17	2

a) *¿Cuál es el promedio del largo de los tornillos de la muestra revisada?*

b) *¿Cuáles son la moda y la mediana del largo de los tornillos en la muestra?*

a. ¿Cómo explicaría la forma de responder las preguntas planteadas en la actividad a niños y niñas de un curso de Educación Básica?

b. Explique, al menos, 2 errores que pueda cometer un alumno al responder las preguntas de la actividad.

Capítulo
V

Probabilidad

Introducción

Desde tiempos muy antiguos, el hombre enfrenta situaciones que quisiera poder anticipar, o prever la forma en que ocurrirán. Esto lo expone a la incerteza inherente a una gran parte de los fenómenos que observamos, sean estos naturales o el resultado de la acción del hombre. Si bien el estudio formal de la incerteza, que comienza alrededor el siglo XVI, aparece ligado específicamente a lo que hoy llamamos *juegos de azar*, la relevancia de la teoría de probabilidades desarrollada trasciende a aplicaciones en ellos. En efecto, la palabra *azar* se aplica a cualquier situación cuyo resultado es incierto, tocando aspectos tanto de la vida diaria (lloverá o no), como de la salud (una persona se enfermará o no), las ciencias sociales (la población estará o no de acuerdo con una medida adoptada por el gobierno), entre muchas otras.

En este capítulo, estudiaremos una forma de cuantificar la incerteza, a través de la probabilidad, presentando una definición formal y resultados que nos ayudan a trabajar con ella. En particular, este capítulo se organiza como sigue: la **Sección 1** muestra la importancia de entender la incerteza como una característica inherente a toda actividad del ser humano, y la necesidad de modelarla. La **Sección 2** muestra la probabilidad como una medida de la incerteza frente a la ocurrencia de ciertos acontecimientos, y entrega su definición frecuentista, para luego generalizarla a partir de sus propiedades. Finalmente, la **Sección 3** discute reglas que permiten obtener probabilidades de acontecimientos complejos, en base a las probabilidades de ocurrencia de fenómenos simples.

1. Motivación

Consideremos, a modo de ejemplo, una situación que enfrentamos desde tiempos remotos: un agricultor está interesado en la cantidad de lluvia que caerá durante cierta temporada de cultivo, para poder anticipar su cosecha. En situaciones como estas, el hombre se ve enfrentado a *incerteza*, en el sentido de que, aun existiendo algunos patrones de comportamiento, le resulta imposible predecir una situación futura con toda seguridad.

Supongamos que, a partir de la observación del comportamiento de las lluvias en temporadas similares de años anteriores, un agricultor ha notado un mínimo de 30 días de lluvia y un máximo de 55 días. Con esta información, ¿puede inferir, con seguridad, que el número de días lluviosos en la temporada siguiente estará entre estas cantidades?

La respuesta es negativa. Si bien la observación de temporadas anteriores entrega información útil acerca de lo que ocurrirá en temporadas futuras, el número de días lluviosos es diferente de un año a otro. Por lo tanto, decimos que existe *variabilidad* en el número de días lluviosos de cada temporada. Esta variabilidad, que se debe tanto al gran número de variables que influyen en los fenómenos meteorológicos, como a la inestabilidad de estos mismos, es la causa de la incerteza a la que nos referimos anteriormente.

Si bien la falta de predictibilidad de acontecimientos futuros debido a su variabilidad sugiere la idea de "desorden", es posible encontrar, en algunas situaciones, ciertos patrones que utilizados de una manera correcta pueden ayudarnos a tomar decisiones. A modo de ejemplo, si un agricultor nota una relación entre la ocurrencia del Fenómeno del Niño y la temporada de lluvias, puede decidir alterar su patrón de siembra, aun cuando no tenga total certeza de que esto volverá a ocurrir en temporadas siguientes.

En la actualidad, conocemos día a día pronósticos meteorológicos para fechas futuras cercanas. Estos pronósticos intentan acotar nuestra incerteza frente a la situación que realmente ocurrirá. Con los avances en el estudio de fenómenos meteorológicos, estas predicciones son bastante certeras. Sin embargo, las situaciones reales observadas posteriormente, pudiendo acercarse al pronóstico, no serán exactamente las predichas debido a la inestabilidad propia de la física de la atmósfera, por lo que aun hay fuentes de incerteza.

A modo de ejemplo, la **Figura V.1** muestra el pronóstico del tiempo para Fiestas Patrias, entre el 14 y el 18 de septiembre de 2013, en la ciudad de Valdivia, entregado el día 13 de septiembre. La figura muestra que la temperatura máxima predicha para el día lunes 16 es de **10 °C**. La **Figura V.2** muestra las condiciones reales observadas el día lunes 16 y el pronóstico para los días siguientes. En la figura observamos que la temperatura máxima observada el día lunes 16 fue de **12 °C**, muy cercana a la predicción de **10 °C**, sin embargo no es exactamente dicho valor.

Valdivia			
Sábado 14	Mín. 0 °C Máx. 12 °C		Nubosidad parcial
Domingo 15	Mín. 3 °C Máx. 11 °C		Nublado y chubascos
Lunes 16	Mín. 4 °C Máx. 10 °C		Nubosidad parcial variando a despejado
Martes 17	Mín. 2 °C Máx. 10 °C		Nubosidad parcial
Miércoles 18	Mín. 1 °C Máx. 11 °C		Nubosidad parcial

Figura V.1: Pronóstico meteorológico entregado el 13 de septiembre de 2013, para los días del 14 al 18 de septiembre del mismo año.

Valdivia			
Lunes 16	Máx. 12 °C		Nubosidad parcial variando a despejado
Martes 17	Mín. 2 °C Máx. 11 °C		Despejado
Miércoles 18	Mín. 1 °C Máx. 13 °C		Nubosidad parcial
Jueves 19	Mín. 6 °C Máx. 10 °C		Nublado y chubascos
Viernes 20	Mín. 5 °C Máx. 11 °C		Nublado y chubascos

Figura V.2: Condiciones meteorológicas observadas el día lunes 16 de septiembre, y el pronóstico para los días del 17 al 20 de septiembre de 2013.

Las **Figuras V.1** y **V.2** también muestran que los pronósticos van siendo actualizados a la luz de nuevos datos. A modo de ejemplo, la **Figura V.1** muestra que el día 13 de septiembre, el pronóstico de temperaturas mínima y máxima para el día 18 de septiembre era de **1 °C** y **11 °C**, respectivamente. La **Figura V.2** muestra que, el día 16 de septiembre, el pronóstico de las mismas temperaturas, mínima y máxima para el día 18 de septiembre era de **1 °C** y **13 °C**, respectivamente. Vemos que este último pronóstico, si bien, muy cercano al anterior, ha sido actualizado a la luz de nueva información recolectada entre el viernes 13 y el lunes 16.

En la actualidad, estudiar, modelar y entender la incerteza es de gran importancia desde distintos puntos de vista. Por una parte, existen con frecuencia situaciones de la vida diaria en las que un ciudadano debiese manejar adecuadamente los conceptos asociados a la incerteza, como: la interpretación de un pronóstico meteorológico, la comprensión de un grado de peligro sísmico, la interpretación de encuestas, de predicciones económicas y de márgenes de error, por mencionar algunas.

A modo de ejemplo, el reporte de la encuesta CEP julio–agosto 2013[1], preparado por el Centro de Estudios Públicos de Chile, afirma en la descripción de la metodología utilizada para el estudio:

"El método de muestreo fue estratificado, aleatorio y probabilístico en cada una de sus tres etapas. …El error muestral se estima en +–3% considerando varianza máxima…".

La información entregada en la afirmación anterior ha sido obtenida a partir de consideraciones probabilísticas y se espera que pueda ser comprendida por la ciudadanía en general.

Por otra parte, y de gran importancia, el estudio de la incerteza en resultados de ciertos experimentos ha permitido avances en las ciencias en la medida en que ha sido posible identificar los patrones que la rigen. Esto ha ocurrido tanto en Física, al modelar las leyes de la termodinámica, como en Medicina, en el estudio del origen de la resistencia bacteriana o en la interpretación de exámenes médicos, en Biología, en el estudio del comportamiento de ciertas especies, en Genética, en el estudio de mutaciones, y en Ciencias Sociales en el estudio de patrones de aprendizaje, obtención de tablas de esperanza de vida, entre otras ciencias y aplicaciones.

[1] Tomado de la página web del Centro de Estudios Públicos de Chile. http://www.cepchile.cl/1_5349/doc/estudio_nacional_de_opinion_publica_julio-agosto_2013.html#.Uj9pUr-Lh1U

2. Cuantificación de la incerteza a través de probabilidades

2.1. Experimento aleatorio

Consideremos un grupo de niños jugando a los dados, donde gana el niño que obtiene el número más alto en un lanzamiento. Antes de realizar los lanzamientos, es imposible predecir quién será el ganador del juego, por lo que nos encontramos ante una situación de incerteza. Para conocer el resultado del juego, cada uno de los niños debe lanzar el dado y registrar el número obtenido.

Consideremos otra situación, donde se desea conocer la preferencia de un grupo de personas frente a la próxima elección presidencial; para eso se tomará una muestra aleatoria simple a partir de este grupo. Esta situación presenta incerteza debido a que no podemos saber de antemano qué personas pertenecerán a la muestra, así como tampoco podemos anticiparnos a las preferencias que estas personas expresarán. Para conocerlas, debemos obtener la muestra, efectuar las preguntas pertinentes y registrar las respuestas.

En las situaciones anteriores, se debe registrar el resultado de un proceso. En el juego de los niños, el proceso consiste en el lanzamiento del dado de cada uno, y el resultado a registrar corresponde al número de la cara superior del dado que obtuvo cada uno de ellos. En la encuesta referida a la próxima elección presidencial, el proceso consiste en la elección de las personas de la muestra y la formulación de la pregunta pertinente, y el resultado corresponde a la preferencia expresada por cada una de ellas. Este tipo de procesos es llamado *experimento aleatorio* y corresponde a cualquier procedimiento o situación que produce un resultado que no es predecible de antemano.

Según la definición anterior, son experimentos aleatorios: observar y registrar la evolución de las notas de un alumno a lo largo del próximo año, u observar y registrar el equipo de fútbol que gana un partido, entre otros ejemplos. En ocasiones como estas, el término "experimento" parece inadecuado, puesto que este sugiere la intervención de un "experimentador" en la generación del resultado a observar. En el ejemplo inicial, los niños deben lanzar los dados, tomando así un rol de "experimentadores". Sin embargo, el clima en un tiempo futuro no requiere de nuestra intervención y seremos meros espectadores de este experimento aleatorio. De todas formas, resulta útil utilizar la palabra "experimento" para algo que produce un resultado, aun cuando no exista un experimentador.

La definición de un experimento ayuda al propósito de explicitar exactamente lo que nos interesa estudiar y qué tipo de resultado nos es relevante. A modo de ejemplo, en el experimento "observar el equipo de fútbol que gana un partido" nos interesa saber el equipo ganador y no, por ejemplo, cuántos goles hizo cada equipo, cuántos cambios de jugadores hubo, si hubo o no tarjetas amarillas, etc.

Según esto, las siguientes descripciones no determinan completamente un experimento aleatorio: esperar un bus, lanzar un dado o invertir en un negocio. Si bien en todas ellas existe aleatoriedad, lo que se ha descrito es únicamente el proceso, pero no se ha especificado qué tipo de resultado se registrará. Sí constituyen descripciones completas de experimentos:

- Esperar un bus y registrar el tiempo de espera hasta que este llegue.
- Esperar un bus y observar el número de personas que vienen en él.
- Lanzar un dado y registrar si se observa un número par o impar en la cara superior.
- Invertir en un negocio y observar el porcentaje de ganancia o pérdida después de un año.

Un aspecto importante en la definición de experimento aleatorio consiste en fijar la información disponible para predecir su resultado. A modo de ejemplo, supongamos que estamos interesados en determinar el número de personas que asistirán a un concierto. El experimento de contar y registrar el número de asistentes, una vez que estos se encuentran ubicados en la sala, presenta aleatoriedad, dado que no conocemos este número de antemano. Sin embargo, los grados de incerteza respecto al número de asistentes pueden ser diferentes para cada persona. El boletero, por ejemplo, conoce el número de entradas vendidas, por lo que la aleatoriedad para él proviene únicamente del hecho de que no puede anticipar cuántas de las personas que adquirieron una entrada tendrán algún inconveniente y no podrán asistir. Sin embargo, una persona común tiene, en general, menos elementos que el boletero para predecir el número de asistentes, por lo que, para ella, existen más fuentes de aleatoriedad y tendrá, por esto, un mayor grado de incerteza.

Del mismo modo, es posible que el controlador de una línea de buses disponga de más elementos para determinar el tiempo de espera de un pasajero en un paradero que este último, dado que el controlador determina la frecuencia de salidas. Pero, aun en este caso, el controlador no podrá predecir el tiempo de espera con exactitud, ya que no controla la congestión vehicular, los tiempos de detención en paraderos, etc. Tanto para el pasajero como para el controlador, el observar y registrar el tiempo de espera corresponde a un experimento aleatorio; sin embargo, ambos poseen diferentes grados de incerteza frente a su resultado, porque ambos manejan distinta información.

Ejercicios

1. Discuta si las siguientes situaciones son o no experimentos aleatorios para la persona que registra su resultado.
   a. Andrés, un niño de séxto básico, pide a un amigo que diga un número par, y registra el resto de dicho número al dividirlo por 2.
   b. Leonor observa y registra el número de llamadas telefónicas que recibe en un día dado.
   c. Usted le pide a un alumno que escoja una letra del abecedario y registra si se obtiene una vocal o una consonante.
   d. El profesor Salinas pasa la lista en su curso y registra el número de alumnos y alumnas asistentes.
   e. Amelia observa a 2 niños que juegan al cachipún, y registra el nombre del niño que gana en cada intento.

2. En cada una de las siguientes situaciones, proponga elementos que las complementen, para poder definirlas, completamente, como un experimento aleatorio.

   a. Una niña hace cálculos con una calculadora.

   b. Andrea mira los edificios en Santiago.

   c. El profesor Soto observa a los niños en el recreo.

   d. La profesora García revisa las pruebas de sus alumnos.

   e. La profesora Pino desarrolla la planificación anual de los contenidos de su curso.

2.2. Grados de posibilidad y niveles de incerteza

Supongamos que estamos interesados en saber si mañana tendremos un día soleado. Si bien no estamos seguros de si esto ocurrirá, en general, estaremos inclinados a pensar que ocurrirá, o no, basados, por ejemplo, en el clima propio del lugar en que nos encontramos, en la estación del año y en el tiempo del día de hoy, entre otros factores.

Por otra parte, si estamos inclinados a pensar que mañana habrá un día soleado, podemos estarlo en mayor o menor medida y, del mismo modo, ocurre si estamos inclinados a pensar que mañana no habrá un día soleado. Puede también ocurrir, por supuesto, que no tengamos una inclinación particular sobre si habrá o no un día soleado.

Según lo anterior, podemos identificar diferentes *grados de posibilidad* de que ocurra cierto resultado, como se muestra en la **Figura V.3**, donde hemos catalogado estos grados desde *Imposible* hasta *Seguro*, moviéndonos a través de grados intermedios. Notemos que, en la figura, la localización de los grados en los extremos, Imposible y Seguro, y del grado central, *Posible*, pueden ser considerados absolutos. Sin embargo, los grados intermedios no corresponden necesariamente a los puntos que se muestran en la escala, sino que a un rango difícil de delimitar.

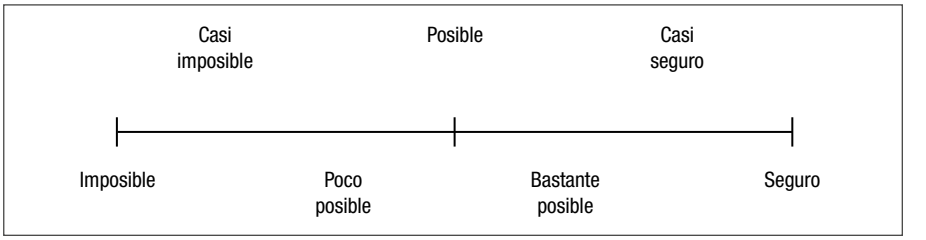


Figura V.3: Diferentes grados de posibilidad de ocurrencia de un resultado.

Cada uno de estos grados de posibilidad implica diferentes niveles de incerteza. En efecto, la certeza es máxima en los extremos de la escala: si calificamos la situación de interés como Imposible, estamos seguros de que ella no ocurrirá y, del mismo modo, si la calificamos como Segura, estamos seguros de que ella ocurrirá. En ambos casos, tenemos completa certeza. Por otra parte, si todo lo que podemos decir es que calificamos la situación de interés como Posible, estaremos ante un grado máximo de incerteza: no podemos inclinarnos ni hacia la ocurrencia ni hacia la no ocurrencia de la situación que nos interesa.

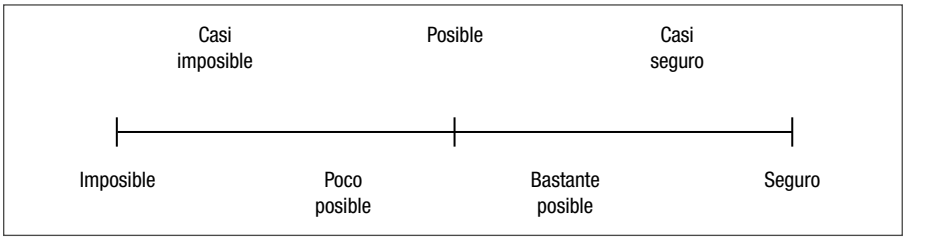
A modo de ejemplo, supongamos que mañana, antes de salir de casa, nos interesa saber si habrá lluvia durante el día. Si sabemos que es Imposible que esto ocurra, tendremos certeza total de que no habrá lluvia y una decisión racional corresponde a no llevar un paraguas. En el otro extremo, si estamos seguros de que lloverá, calificaremos esta situación como Segura, tendremos certeza de que habrá lluvia y una decisión racional corresponde a llevar el paraguas. Sin embargo, si calificamos la situación como *Poco posible*, nuestra incerteza crece, puesto que, si bien estamos inclinados a pensar que no lloverá, no tenemos certeza sobre esto, y nuestra decisión se torna más difícil. En el centro de la escala, si calificamos la situación como Posible, estamos ante incerteza máxima: no estamos inclinados hacia uno ni otro lado, y nuestra decisión sobre llevar o no paraguas se torna aun más difícil.

En resumen

- Los *grados de posibilidad*: constituyen una escala cualitativa de las oportunidades de que ocurra alguna situación. Sus categorías van desde *Imposible* hasta *Seguro*.
- Cada grado de posibilidad implica un nivel diferente de incerteza.

Ejercicios

1. Considere las siguientes situaciones e indique el grado de posibilidad que usted le asignaría a su ocurrencia. Indique los supuestos que realiza para evaluar dicho grado. Puede utilizar la escala:



   a. Al contar las flores de un jardín, que usted encuentre exactamente 135 flores.
   b. Al medir la altura de un niño de tercero básico, que este mida entre 1,20m y 1,30m.
   c. Que mañana caiga un meteorito en alguna zona del país.
   d. Que Andrés gane el premio máximo del Kino al jugar un solo cartón. Considere que este juego consiste en elegir 15 números, sin repetición, entre el 1 y el 25, ambos números incluidos. El(los) cartón(es) ganador(es) será(n) aquél(los) cuyos números coincidan con los 15 números, sin repetición, extraídos desde una tómbola que contiene los números del 1 al 25, inclusive.

2. Para cada una de las situaciones en el ejercicio anterior, discuta si le parecería razonable que diferentes personas asignaran grados de posibilidad diferentes a los suyos. Explique por qué.

2.3. Noción de probabilidad

Notamos que la escala de posibilidad de ocurrencia de un resultado, mostrada en la **Figura V.3**, corresponde a una medida cualitativa, no numérica, para la posibilidad de ocurrencia de un suceso o situación. En esta sección y las siguientes trataremos el concepto de *probabilidad* de ocurrencia de una situación, la que corresponde a una medida cuantitativa de la posibilidad de ocurrencia de esta.

Como primera noción, la probabilidad de ocurrencia de una situación corresponderá a un valor ente **0** y **1**, que también pueden expresarse como **0%** y **100%** de modo que una situación catalogada como Imposible tendrá probabilidad de ocurrencia **0** y, en el otro extremo, una situación catalogada como Segura tendrá probabilidad de ocurrencia **1**. En la medida en que una situación tiene mayor posibilidad de ocurrir, su probabilidad de ocurrencia se acercará a **1**. Cuando la situación es catalogada como Posible, su probabilidad de ocurrencia será de **0,5**. Esta situación se ilustra en la **Figura V.4**, que muestra la correspondencia entre grados de posibilidad y valores de probabilidad, ambos asociados a la ocurrencia de una misma situación.

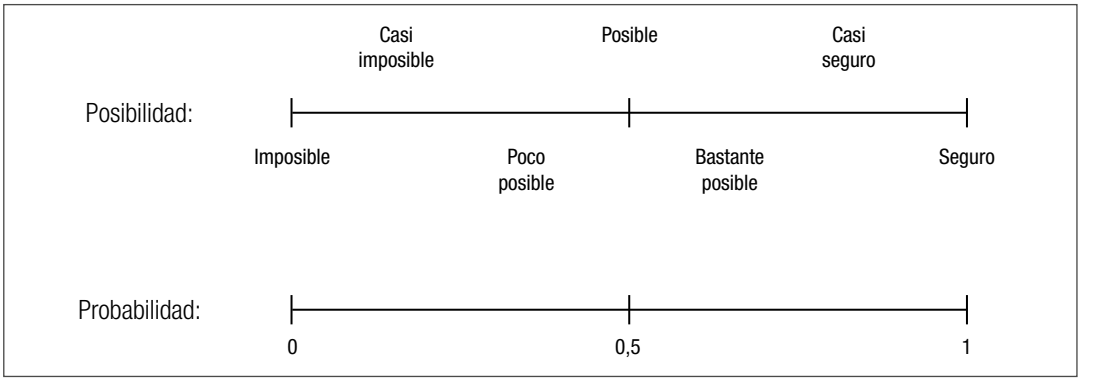


Figura V.4: Superior: Diferentes grados de posibilidad de ocurrencia de un resultado. Inferior: Probabilidad de ocurrencia del mismo resultado.

A modo de ejemplo, si nos dicen que la probabilidad de que a un niño, tomado al azar entre los alumnos del curso, le guste el fútbol es **0,9**, es decir, cercano a **1** en la escala de probabilidad en la **Figura V.4** inferior, podríamos decir que estamos *Casi Seguros* de que una vez elegido el niño, a este le gustará el fútbol. Esto se representa cercano al extremo derecho de la escala de posibilidad en la **Figura V.4** superior. Lo contrario ocurre si la misma probabilidad es **0,1**, cercano al extremo izquierdo tanto en la escala de posibilidad como en la de probabilidad. Por otra parte, si nos dicen que la probabilidad de que al niño elegido le guste el fútbol es **0,5**, podríamos decir que es igualmente posible o imposible que a este le guste el fútbol.

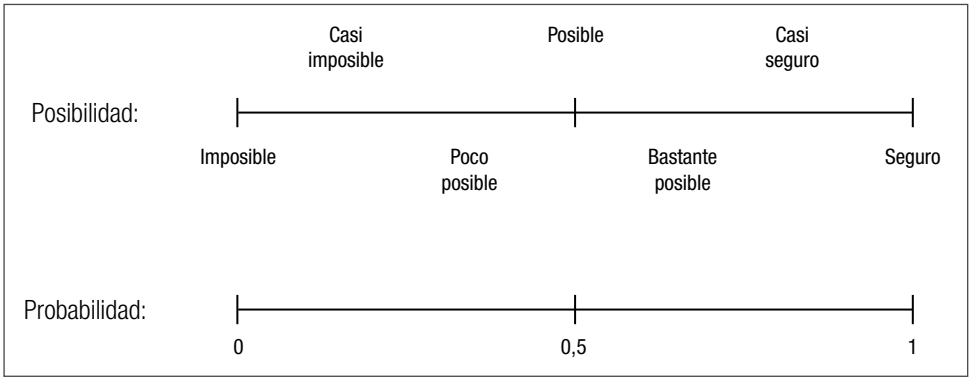
En la sección que sigue, daremos una definición formal de probabilidad.

En resumen

Probabilidad: corresponde a una medida cuantitativa de las posibilidades de que ocurra una situación. Sus valores están entre **0** y **1**.

Ejercicios

1. Considere las siguientes situaciones y sus probabilidades. En cada una de ellas, indique el grado de posibilidad de ocurrencia que usted asignaría, según el valor de la probabilidad dada. Puede utilizar la figura:



   a. La probabilidad de que al lanzar una moneda equilibrada salga cara es **0,5**.
   b. La probabilidad de que un profesor revise el libro de clases un día domingo es **0**.
   c. La probabilidad de que, en un día dado, un alumno tenga clases de matemática es **1**.
   d. La probabilidad de que un niño coma las legumbres de su almuerzo está entre **0,3** y **0,7**.
   e. La probabilidad de que un profesor cite a reunión de apoderados en el mes de marzo es mayor que **0,85**.
   f. La probabilidad de que, en un curso formado por **30** niños, al menos **2** tengan cumpleaños el mismo día es mayor que **0,5**.

2. Considere las siguientes situaciones y sus posibilidades. En cada una de ellas, indique aproximadamente el valor o los valores posibles de probabilidad que se le pueden asociar. Puede usar la figura del ejercicio anterior.

   a. Estoy seguro que mañana va a llover.
   b. Es muy poco posible que pase una micro vacía a esta hora.
   c. Es bastante seguro que a esta hora nos encontraremos con congestión en la carretera.
   d. Estoy casi seguro de que estos serán los números de la lotería.
   e. No tengo idea de la posibilidad de que Chile gane algún mundial.
   f. Estoy seguro que el próximo candidato a presidente hará promesas de campaña que no cumplirá.
   g. No se sabe cuándo ni dónde será el próximo terremoto.
   h. Estoy seguro que si hago ejercicios tendré una mejor salud.

2.4. Definición frecuentista de probabilidad

Teniendo ya la noción de lo que representan distintos valores de una probabilidad, la definiremos formalmente, haciendo un primer acercamiento a esta definición a través de diferentes ejemplos.

Consideremos el experimento de lanzar una moneda y registrar el lado que se observa tras caer. Nos interesa dar una definición para la probabilidad de observar una cara. Una manera de medir qué tan probable es observar una cara corresponde a repetir muchas veces el experimento, de manera independiente y manteniendo cada vez las mismas condiciones, y registrar en cuántas oportunidades se obtiene una cara. A modo de ejemplo, supongamos que lanzamos 10 veces la moneda de interés y registramos el resultado en cada lanzamiento, obteniendo la siguiente secuencia, donde anotamos C si se observa una cara, y S si se observa un sello:

C C S S C S S S C S

Vemos que, en estos 10 lanzamientos, se obtuvo 4 caras, es decir, la frecuencia relativa de caras fue de $\frac{4}{10} = 0{,}4$, que corresponde a una frecuencia relativa porcentual de 40%, lo cual nos da una idea de qué tan posible es observar una cara. Sin embargo, sabemos que si repitiéramos los 10 lanzamientos los resultados serían diferentes. A modo de ejemplo, al lanzar la misma moneda otras 10 veces obtenemos la secuencia:

S S C S S C C C C S

En ella se observan 5 caras, es decir, una frecuencia relativa de 0,5, o relativa porcentual de 50%. Si unimos los resultados de los 20 lanzamientos, observamos un total de 9 caras, por lo que la frecuencia relativa de caras es $\frac{9}{20} = 0{,}45$, o relativa porcentual 45%. La Tabla V.1 muestra los resultados obtenidos al lanzar la moneda 10, 20, 100, 500, 1.000 y 2.000 veces consecutivas.

Número de lanzamientos	Número de caras obtenidas	Frecuencia relativa de caras obtenidas
10	4	$\frac{4}{10} = 0{,}40$
20	9	$\frac{9}{20} = 0{,}45$
100	44	$\frac{44}{100} = 0{,}44$
500	249	$\frac{249}{500} = 0{,}498$
1.000	493	$\frac{493}{1.000} = 0{,}493$
2.000	1.003	$\frac{1.003}{2.000} = 0{,}5015$

Tabla V.1: Frecuencia relativa de caras obtenidas en diferentes números de lanzamientos de una moneda.

La **Figura V.5** muestra el comportamiento de las frecuencias relativas en mayor detalle, en la medida que lanzamos repetidamente la moneda. En el eje horizontal, o de las abscisas, la figura indica el número de lanzamientos realizados, mientras que en el eje vertical, o de las ordenadas, se muestra la frecuencia relativa de caras obtenidas hasta el lanzamiento indicado. La figura muestra cómo esta frecuencia relativa se estabiliza al crecer el número de lanzamientos, es decir, cómo se acerca más y más a cierto valor. Este valor es el que definimos como la *probabilidad de obtener una cara al lanzar una vez una moneda*. En la figura, vemos que el valor al que se acerca cada vez más se parece a **0,5**, que, por simetría de la moneda, es el valor que esperaríamos si es que la moneda no está cargada hacia uno u otro lado.

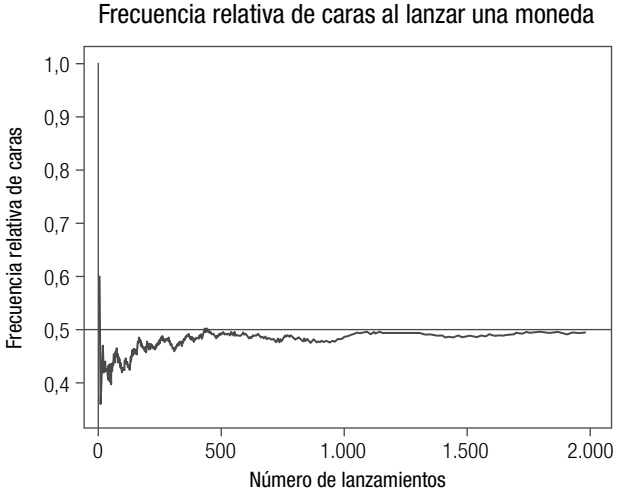


Figura V.5: Frecuencia relativa de caras para diferentes números de lanzamientos de una moneda.

Consideremos, ahora, el experimento de extraer una bolita desde una urna que contiene bolitas blancas y negras en diferentes proporciones, registrar su color y devolver luego la bolita a la urna. Nos interesa conocer los valores de las probabilidades de obtener una bolita blanca y de obtener una bolita negra. La **Figura V.6** muestra el resultado de **25** repeticiones de este experimento, donde a la derecha de la urna se muestran las bolitas que han sido extraídas.

En la línea inferior de bolitas, moviéndonos de izquierda a derecha, vemos que la primera bolita extraída fue negra, con lo que las frecuencias relativas de bolitas blancas y negras, hasta ese momento, son **0** y **1**, respectivamente. La bolita es repuesta luego en la urna, de modo de que, al extraer una segunda bolita, la configuración de bolitas en la urna sea idéntica a la inicial. Dado que la segunda bolita extraída también es negra, las frecuencias relativas de bolitas blancas y negras continúan siendo **0** y **1**, respectivamente. Ya en la tercera repetición se obtiene la primera bolita blanca, con lo que la frecuencia relativa de bolitas blancas sube a $\frac{1}{3}$, aproximadamente **0,33**, mientras que la frecuencia relativa de bolitas negras baja a $\frac{2}{3}$, aproximadamente **0,67**.

Repitiendo el experimento, se observa que, en las **25** extracciones con reposición que se muestran en la **Figura V.6**, se obtuvieron **7** bolitas blancas y **18** bolitas negras, con lo que las frecuencias relativas de bolitas blancas y de bolitas negras, al finalizar las repeticiones del experimento, son $\frac{7}{25} = 0{,}28$, y $\frac{18}{25} = 0{,}72$, respectivamente.

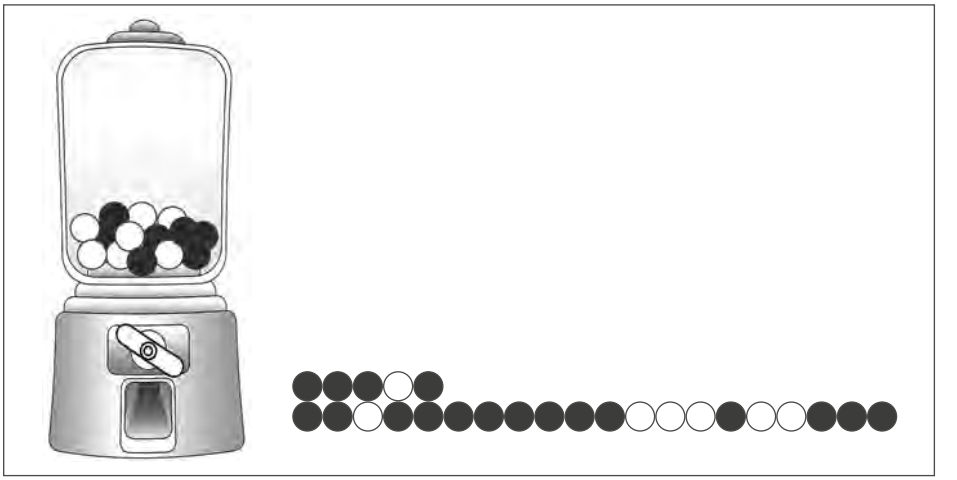


Figura V.6: 25 extracciones de una bolita a partir de la urna, con reposición de la bolita después de cada extracción.

Para pensar

Notemos que las frecuencias relativas obtenidas, de bolitas blancas y de bolitas negras, siempre suman 1 (a modo de ejemplo, después de 25 extracciones, 0,28 + 0,72 = 1). ¿Siempre se cumplirá esto? ¿por qué?

¿Qué implica esto sobre las probabilidades que buscamos, de obtener una bolita blanca y de obtener una bolita negra? ¿Es necesario conocer ambas probabilidades o basta con conocer solo una de ellas?

La **Figura V.7** muestra la evolución de las frecuencias relativas de bolitas blancas y negras, a través de las 25 repeticiones del experimento. En ella, se observa aun bastante variabilidad de una extracción a la siguiente.

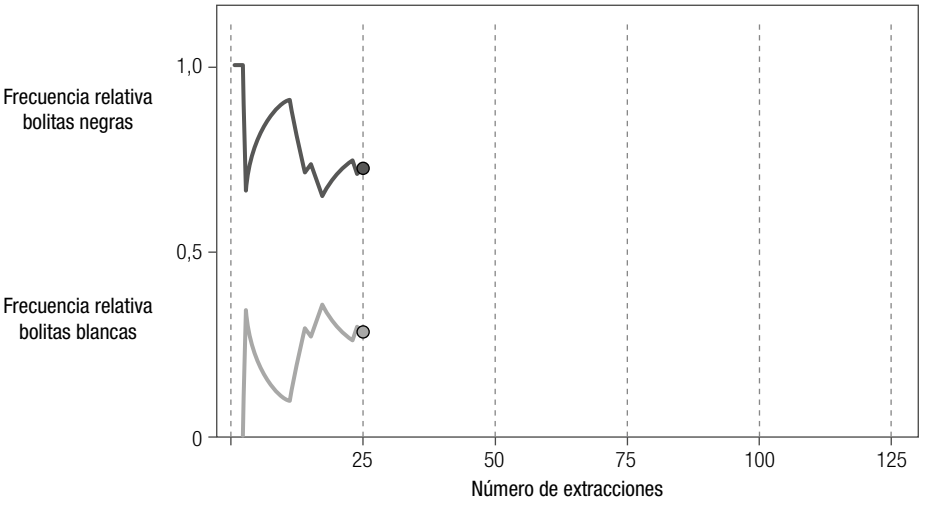


Figura V.7: Evolución de las frecuencias relativas de bolitas blancas y negras, a través de 25 extracciones sucesivas con reposición.

La **Figura V.8** muestra, ahora, el resultado de **120** extracciones de bolitas a partir de la misma urna, reponiendo cada bolita después de extraerla. Contando los números de bolitas blancas y negras extraídas, encontramos que, al finalizar las **120** extracciones, las frecuencias relativas de interés corresponden a $\frac{32}{120}$, aproximadamente **0,27**, y $\frac{88}{120}$, aproximadamente **0,73**. La **Figura V.9** muestra la evolución de las frecuencias relativas de bolitas blancas y negras, a través de las **120** repeticiones del experimento. En ella, se observa que la variabilidad de estas frecuencias decrece en la medida que crece el número de extracciones.

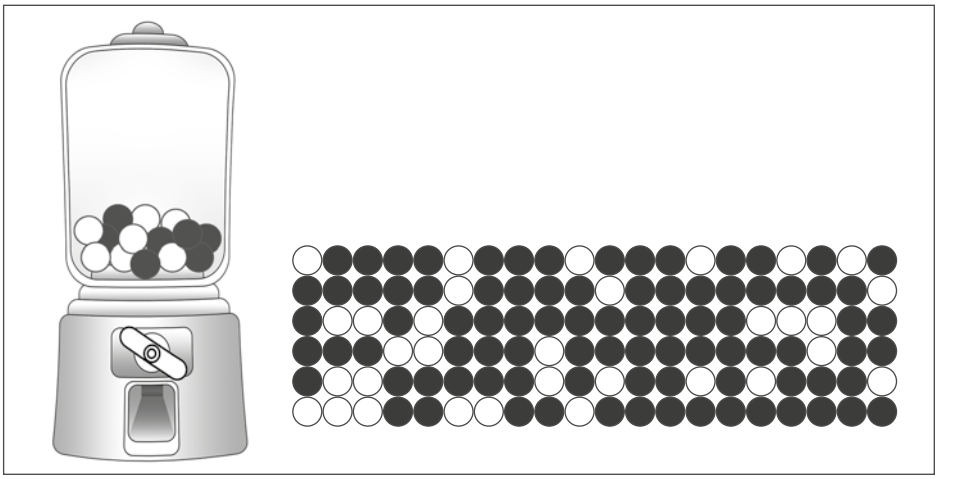


Figura V.8: 120 extracciones de una bolita a partir de la urna, con reposición de la bolita después de cada extracción.

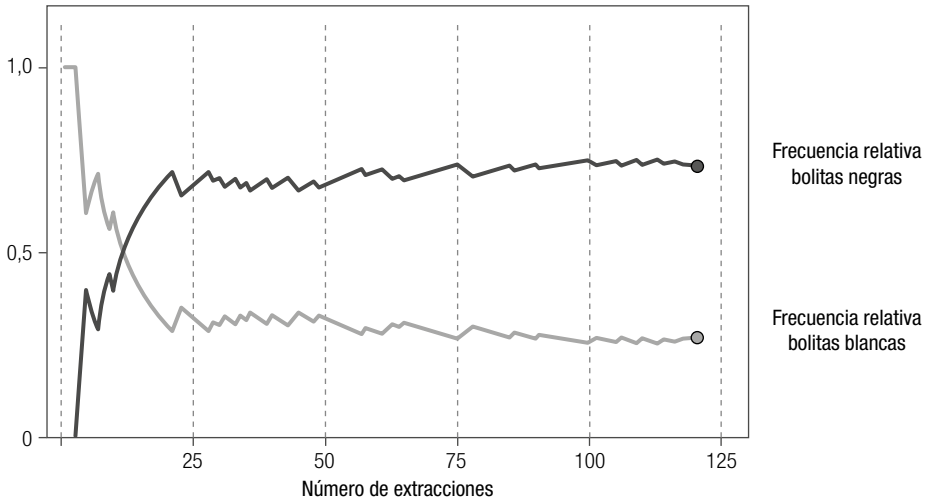


Figura V.9: Evolución de las frecuencias relativas de bolitas blancas y negras obtenidas después de 120 extracciones sucesivas.

Finalmente, la **Figura V.10** muestra la evolución de las frecuencias relativas de bolitas blancas y negras a través de **3.000** repeticiones del experimento. Se observa ahora una gran estabilidad de estas frecuencias. Del gráfico leemos que las frecuencias relativas de bolitas blancas y negras se acercan, aproximadamente, a **0,25** y **0,75**, respectivamente. Se puede mostrar que estos valores corresponden exactamente a las proporciones de bolitas blancas y negras en la urna: **7** bolitas de color blanco y **21** bolitas de color negro. Estudiaremos este punto en secciones siguientes.

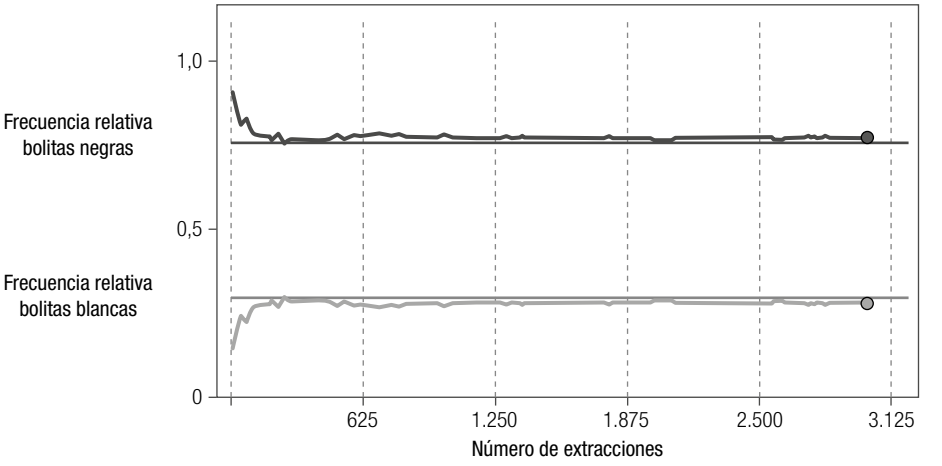


Figura V.10: Evolución de las frecuencias relativas de bolitas blancas y negras, después de 3.000 extracciones sucesivas.

En general, se puede demostrar formalmente que siempre, al repetir un gran número de veces un experimento asegurando que sus repeticiones sean realizadas de manera *independiente* y bajo exactamente las *mismas condiciones*, la frecuencia relativa de ocurrencia de un mismo resultado tiende, o se acerca cada vez más, a un valor, que es el que llamamos *probabilidad de ocurrencia* de este resultado.

Existe una definición probabilística formal del concepto de *independencia* entre **2** situaciones. Informalmente, se dice que **2** situaciones son independientes si la ocurrencia de una de ellas no afecta la probabilidad de ocurrencia de la otra. Otra manera de entenderlo es identificar situaciones donde, por el contrario, los experimentos no son independientes. Para ilustrarlo, supongamos que en un grupo de personas, el experimento consiste en pedirle a una de ellas que exprese en voz alta su postura frente a un tema delicado de actualidad. Dicha persona indicará su postura, la que será escuchada por el resto del grupo. ¿Cree usted que esto pudiese influenciar la respuesta de una segunda persona cuando se le pregunta su opinión sobre el mismo tema? A modo de ejemplo, ¿la influenciaría si de esto dependiese su potencial contratación en un trabajo? Es posible que la segunda persona entregue una respuesta influenciada, de alguna manera, por la respuesta de la primera persona. En una situación como esta, las repeticiones del experimento consisten en la realización de la pregunta a cada una de las personas del grupo, sin embargo, probablemente, sus respuestas no serán independientes.

Para pensar

¿Cómo cree usted que pudiese modificarse el experimento de manera que repeticiones de este fuesen independientes?

Una segunda condición enunciada para la estabilización de la frecuencia relativa de una situación corresponde a que las repeticiones del experimento deben ocurrir bajo las *mismas condiciones*. A modo de ejemplo, si el experimento corresponde a observar y registrar la distancia a la que cae un avión de papel tras lanzarse, repeticiones debiesen considerar al mismo lanzador, las mismas condiciones de viento, y aviones fabricados con el mismo tipo de papel y según el mismo modelo, en caso de utilizar aviones diferentes.

En resumen

- Al repetir un número suficientemente grande de veces un experimento, de manera independiente y en las mismas condiciones, las frecuencias relativas de ocurrencia de una situación dada se estabilizan, tendiendo o acercándose cada vez más a cierto valor.
- Según la definición frecuentista, este valor corresponde a la *probabilidad de ocurrencia* de dicho resultado.

Ejercicios

1. De acuerdo a la definición frecuentista de probabilidad, ¿qué probabilidades aproximadas cree usted que corresponden a cada una de las siguientes situaciones? Justifique.

   a. Que al lanzar 3 monedas, se obtengan exactamente 2 caras.
   b. Que al lanzar 2 veces un dado, la suma de las caras superiores obtenidas sea igual a 5.
   c. Que al lanzar 2 veces un dado, el número obtenido en la cara superior en el primer lanzamiento sea menor que el número obtenido en el segundo.
   d. Que al extraer una bolita de una caja que tiene 5 bolitas negras y 2 bolitas blancas, se obtenga una bolita de color negro.

2. En el ejercicio anterior, diseñe y realice experimentos para verificar cuán exactamente se cumplen sus conjeturas sobre el valor de las probabilidades.

3. Suponga que la probabilidad de que al nacer un bebé este sea varón es 0,5. A lo largo de un año completo, ¿en cuál de los siguientes lugares cree usted que habrá más días en los cuales más del 60% de los bebés nacidos sean varones? Justifique en términos de la variabilidad de las frecuencias relativas.

   a. En un hospital grande, con alrededor de 100 nacimientos por día.
   b. En un hospital pequeño, con alrededor de 10 nacimientos por día.
   c. No habrá diferencia entre estos 2 lugares.

4. Para desarrollar este ejercicio, usted requiere de un dado equilibrado de **6** caras. Considere el experimento de lanzar este dado **2** veces, y realícelo de manera sucesiva, registrando sus resultados en la siguiente tabla. Puede guiarse por los ejemplos dados para las **3** primeras repeticiones que se muestran en ella. Si lo estima pertinente, puede no considerar los resultados dados como ejemplo, y utilizar únicamente los resultados de los lanzamientos realizados por usted.

Número de repetición	Resultado del primer dado	Resultado del segundo dado	¿Es el número obtenido en el primer dado menor que el número obtenido en el segundo?	Número acumulado de experimentos en que el número obtenido en el primer dado es menor que el número obtenido en el segundo	Frecuencia relativa de veces en que el número obtenido en el primer dado es menor que el número obtenido en el segundo
1	2	3	Sí	1	$\frac{1}{1} = 1$
2	4	4	No	1	$\frac{1}{2} = 0{,}5$
3	1	6	Sí	2	$\frac{2}{3} \approx 0{,}67$
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					

a. ¿Observa usted que la frecuencia relativa se acerca cada vez más a un valor determinado?
b. De acuerdo a esto, ¿cuál cree usted que es aproximadamente la probabilidad de que, al lanzar 2 veces un dado, el número obtenido en el primer lanzamiento sea menor que el número obtenido en el segundo?
c. Grafique las frecuencias relativas que registró en la tabla. En el eje de las abscisas, u horizontal, represente el número de repetición del experimento, entre 1 y 22, y en el eje de las ordenadas, o vertical, represente la frecuencia relativa asociada a dicha repetición, que registró en la columna a la derecha de la tabla. Puede guiarse por las Figuras V.5 o V.7. ¿Qué ocurre con la variabilidad de la frecuencia relativa en la medida que aumentamos el número de experimentos realizados?

5. Considere las siguientes situaciones y discuta, para cada una de ellas, la factibilidad de realizar repeticiones de un experimento para aproximar su probabilidad de ocurrencia.
   a. Al dirigirse al banco a realizar un trámite, que usted encuentre una fila vacía.
   b. Que mañana ocurra un día soleado en Temuco.
   c. Que en cierto lugar, ocurran al menos 5 terremotos cada 100 años.
   d. Que cuando un niño pregunte a su mamá si puede ir al concierto de su grupo musical favorito, ella le responda de manera afirmativa.
   e. Que al elegir un niño de manera aleatoria en un curso, su sabor favorito de helado sea el de vainilla.

6. Al lanzar 500 veces una moneda, se han obtenido 398 caras y 102 sellos.
   a. ¿Cree que la moneda está cargada? Justifique.
   b. ¿Cuál cree usted que es, aproximadamente, la probabilidad de que salga cara en dicha moneda?

3. Asignación de probabilidades

3.1. Espacio muestral de un experimento

Al definir un experimento, hicimos notar la necesidad de explicitar claramente la información que se registrará a partir de la observación de un fenómeno. A modo de ejemplo, dijimos que "esperar a que pase un bus" no es suficiente para describir un experimento, sino que se debió agregar: "y registrar el tiempo que demoró el bus en pasar, desde que el observador llegó al paradero". El siguiente paso para modelar una situación de incerteza, una vez descrito el experimento, corresponde a identificar cuáles son sus posibles resultados.

Por ejemplo, si lanzamos una moneda y registramos el lado de la moneda que queda visible, los resultados posibles son "cara" y "sello"; si lanzamos un dado de 6 caras y observamos el número en la cara superior, los resultados posibles son los enteros entre 1 y 6, si registramos las condiciones de nubosidad mañana del mediodía, podemos registrar los resultados posibles como "nublado", "parcialmente nublado" y "despejado" (aunque no es la única forma de hacerlo). En cada uno de los experimentos que hemos mencionado, el conjunto de todos los resultados posibles descritos se denomina *espacio muestral*. La **Figura V.11** muestra, a modo de ejemplo, un espacio muestral posible para el experimento de observar la nubosidad de mañana al mediodía.

Figura V.11: Al observar la nubosidad de mañana al mediodía, podemos definir el espacio muestral como el conjunto de los resultados "nublado", "parcialmente nublado" y "despejado".

Algunas veces, cada elemento del espacio muestral puede corresponder a una combinación de resultados intermedios, como sucede si lanzamos 3 veces una moneda y registramos la secuencia que se observa. En este caso, el espacio muestral corresponde a los resultados:

CCC CCS CSC SCC CSS SCS SSC SSS

Donde, a modo de ejemplo, "CCS" indica que se obtuvo cara en los dos primeros lanzamientos y sello en el tercero.

Al describir un espacio muestral se debe asegurar que:

- Se incluyan todos los resultados posibles.
- No existan superposiciones entre ellos.

En el primer caso, consideremos el experimento "observar y registrar la nubosidad de mañana al mediodía". Los resultados "nublado" y "despejado" no constituyen un espacio muestral adecuado, puesto que puede ocurrir que las condiciones sean de nubosidad parcial. En este caso, el resultado del experimento se encontraría fuera del espacio muestral descrito.

En el segundo caso, consideremos el experimento de lanzar un dado. Los resultados: impar, 2, 4, 5 y 6, no constituyen un espacio muestral adecuado pues, en caso de obtener un 5, este resultado sería a la vez impar, correspondiendo entonces a 2 resultados del espacio muestral.

Según lo anterior, podemos decir que una vez realizado el experimento, su resultado debe ser *un y solo un elemento del espacio muestral.*

En resumen

El *espacio muestral* corresponde a todos los posibles resultados del experimento.

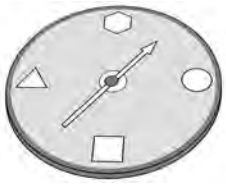
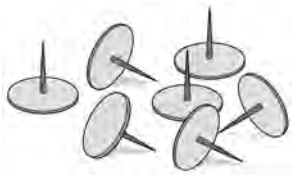
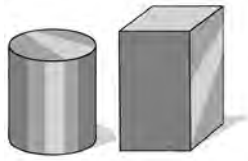
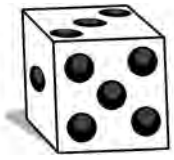
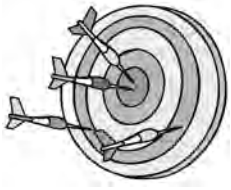
Al describir un espacio muestral, se debe asegurar que:

- Se incluyan todos los resultados posibles.
- No existan superposiciones entre ellos.

Ejercicios

1. Describa el espacio muestral en cada uno de los siguientes experimentos aleatorios.
   a. Leonor observa y registra el número de llamadas telefónicas que recibe en un día dado.
   b. Sebastián lanza un avión de papel y mide la distancia desde donde él está hasta el lugar en que cayó el avión, con una huincha de medir cuya mayor precisión son los centímetros.
   c. Usted le pide a un alumno que escoja una letra del abecedario y registra si se obtiene una vocal o una consonante.
   d. El profesor Rojas pasa lista en su curso y registra el número de alumnos asistentes.

2. Suponga que usted debe preparar una actividad con sus alumnos, para trabajar con el concepto de espacio muestral. Considere cada uno de los siguientes materiales concretos:

 Ruleta	 Chinches	 Cuerpos geométricos
 Dado	 Tiro al blanco	 Lápices de diferentes colores

   a. Proponga un experimento aleatorio que pueda desarrollar con cada uno de los materiales.
   b. Para cada uno de los experimentos propuestos, describa un posible espacio muestral.

3.2. Sucesos o eventos y su ocurrencia

En secciones anteriores, hemos discutido la asignación de probabilidades a ciertas situaciones, como, por ejemplo, obtener una cara al lanzar una moneda, u obtener una bolita negra al extraer bolitas desde una urna. En esta sección, formalizaremos lo que se entiende por situación, en términos de un espacio muestral dado. Estas situaciones serán llamadas *sucesos* o *eventos*.

Para ilustrar, consideremos un grupo de niños jugando a extraer una ficha desde una urna, donde cada una de las fichas muestra uno de 4 símbolos posibles: círculo, triángulo, cruz o rombo, los que se muestran en la Figura V.12.

Figura V.12: Cuatro posibles símbolos indicados en las fichas contenidas en una urna.

Los niños expresan las siguientes inquietudes:

- Pedro postula que en la urna hay más círculos que cualquiera de los símbolos restantes por separado, por lo que le interesa conocer la probabilidad de que, al extraer una ficha desde la urna, esta muestre un círculo.
- Antonia postula que en la urna hay menos cruces que cada uno de los símbolos restantes por separado, por lo que le interesa conocer la probabilidad de que, al extraer una ficha desde la urna, esta muestre una cruz.
- Felipe y Diego deciden que, al extraer una ficha desde la urna, ganará Felipe si se obtiene un triángulo o una cruz, y ganará Diego en caso contrario. A Felipe le interesa conocer su probabilidad de ganar el juego, es decir, de obtener un triángulo o una cruz.

En los ejemplos anteriores, notamos que nos hemos referido a la probabilidad de ocurrencia de:

- Un único elemento del espacio muestral por separado: círculo, en el caso de Pedro, y cruz, en el caso de Antonia.
- Varios elementos del espacio muestral: triángulo o cruz, en el caso de Felipe y Diego.

Notamos entonces que las situaciones a las que estamos asignando probabilidades corresponden a uno o más elementos del espacio muestral. Cada una de estas situaciones se denomina *suceso* o *evento*. A modo de ejemplo, consideremos el experimento de lanzar 2 dados y registrar la suma de sus caras superiores. En este caso, el espacio muestral corresponde a los resultados 2, 3, 4, ..., 11 y 12. Algunos sucesos o eventos pueden ser las situaciones: "el resultado es múltiplo de 4", que corresponde al grupo de resultados 4, 8 y 12, o a "el resultado es un número de 2 cifras" que corresponde al grupo de resultados 10, 11 y 12. Además notemos que el espacio muestral completo también corresponde a un suceso o evento, dado que implica un grupo particular de resultados posibles.

En resumen

- Dado un experimento y su espacio muestral, un *suceso* o *evento* corresponde a un grupo de resultados de este espacio.
- Es posible asignar una probabilidad a la ocurrencia de cada uno de estos sucesos o eventos.

Ahora, debemos establecer lo que significa que un suceso o evento ocurra. Para ello, consideremos nuevamente el juego de Felipe y Diego, y supongamos que estamos interesados en conocer la probabilidad de que gane Felipe. El suceso de interés corresponde al par formado por los símbolos triángulo y cruz, y el suceso ocurrirá cuando, al extraer una ficha de la urna, se obtenga cualquiera de estos 2 resultados, es decir, un triángulo o una cruz, dado que en cualquiera de los 2 casos ganará Felipe.

Siguiendo la idea anterior, en general, decimos que un *suceso* o *evento ocurre* cuando el resultado observado del experimento corresponde a uno de los resultados que lo forman.

Para pensar

¿Por qué el espacio muestral corresponde a un suceso "seguro", en el sentido de que siempre ocurre?

Hacia el lado inverso, decimos que un resultado es *favorable a un suceso dado* cuando forma parte de este. Unido a lo que hemos señalado sobre la ocurrencia de sucesos, podemos decir que un resultado es favorable a un suceso, si observar dicho resultado implica necesariamente la ocurrencia del suceso. A modo de ejemplo, si consideramos el lanzamiento de 2 dados y el suceso "la suma obtenida es un número de 2 cifras", decimos, por ejemplo, que la suma 10 es favorable a dicho evento, dado que, si al realizar el experimento la suma obtenida es 10, necesariamente el suceso "la suma obtenida es un número de 2 cifras" ha ocurrido. Lo mismo ocurre con los resultados 11 y 12.

En resumen

- Un s*uceso* o *evento ocurre* cuando el resultado del experimento es uno de los resultados que lo forman.
- Un resultado es *favorable a un suceso o evento* cuando observar este resultado implica la ocurrencia del suceso o evento.

Por otra parte, consideremos un experimento cualquiera y 2 sucesos definidos sobre su espacio muestral.

Para pensar

¿Es posible que ambos sucesos ocurran simultáneamente?

En el experimento que seguimos, sobre la extracción de fichas desde una urna, consideremos los sucesos: "el símbolo en la ficha extraída es un polígono" y "el símbolo en la ficha extraída está formado por entre 1 y 3 segmentos". El primer suceso ocurre si el símbolo en la ficha extraída es triángulo o rombo, mientras que el segundo suceso ocurre si el símbolo en la ficha extraída es triángulo o cruz. De este modo, en caso de que el resultado del experimento fuese un triángulo, ambos sucesos habrán ocurrido de manera simultánea. Esto se debe a que el triángulo es un resultado que forma parte de ambos sucesos.

En otro ejemplo, al lanzar 2 dados y registrar su suma, consideremos los eventos: "la suma obtenida es un número par" y "la suma obtenida es un número de 2 cifras". Los resultados en el primer suceso son:

2, 4, 6, 8, 10 y 12

Mientras que los resultados en el segundo suceso son:

10, 11 y 12

Para que estos 2 sucesos ocurran de manera simultánea, la suma observada debe ser 10 o 12, es decir, los 2 resultados del espacio muestral que pertenecen a ambos sucesos. De este modo, 2 sucesos cualesquiera pueden ocurrir de manera simultánea solo si tienen, al menos, un resultado en común.

En resumen

- 2 sucesos dados *ocurren simultáneamente* si el resultado observado del experimento es parte de ambos.
- Para que esto ocurra, los sucesos deben tener, al menos, un resultado en común.

Ejercicios

1. Para cada uno de los siguientes experimentos describa un suceso o evento que contenga un solo elemento, y uno que contenga más de uno. Exprese estos sucesos en palabras e indique sus elementos.

   a. Leonor observa y registra el número de llamadas telefónicas que recibe en un día dado.

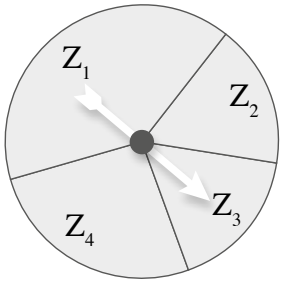
b. Sebastián lanza un avión de papel y mide la distancia desde donde él está hasta el lugar en que cayó el avión. Sebastián realiza la medición con una huincha cuya mayor precisión son centímetros.

c. Usted pide a un alumno que escoja una letra del abecedario y registra si se obtiene una vocal o una consonante.

d. El profesor Rojas pasa lista en su curso y registra el número de alumnos asistentes.

2. Para cada uno de los siguientes experimentos y sucesos de interés, identifique los resultados favorables y no favorables a estos últimos.

a. Al girar la siguiente ruleta, donde las zonas Z_2 y Z_3 tienen el mismo ángulo del centro, interesan los sucesos: "la flecha queda en una de las zonas de menor área" y "la fecha queda en la zona de mayor área".



b. Al extraer una moneda de una caja que contiene 5 monedas de $500, 4 de $100, 7 de $50, 3 de $10, 11 de $5 y 4 de $1, interesan los sucesos "el valor de la moneda es menor que $1.000", "el valor de la moneda es mayor que $5", y "el valor de la moneda está entre $500 y $1, ambos valores incluidos".

3.3. Propiedades que se desprenden de la definición frecuentista de probabilidad

La definición frecuentista que hemos dado al concepto de probabilidad determina ciertas propiedades que deben cumplir todas las probabilidades asociadas a sucesos de un mismo experimento. Revisaremos estas propiedades a continuación.

3.3.1 Una probabilidad es un valor entre 0 y 1

Retomemos el experimento que seguimos en la **Sección 2.4**, de extraer una bolita a partir de una urna que contiene bolitas blancas y negras, y registrar su color. Estudiamos allí las secuencias de frecuencias relativas de ocurrencia de bolitas blancas y negras y, en una medición realizada, se obtuvo que después de **25** extracciones, las frecuencias de ocurrencia de bolitas blancas y negras eran de **0,28** y **0,72**, respectivamente, moviéndose luego a **0,27** y **0,73**, después de **120** extracciones, y a valores cercanos a estos, al continuar repitiendo el experimento, como se muestra en la **Figura V.10**. Tal como ocurre con estos valores, la definición de frecuencia relativa asegura que ellos siempre serán mayores o iguales a **0**, y menores o iguales a **1**. Luego, los valores a los que ellos se acercan cada vez más también deben serlo, de donde se deduce que toda probabilidad debe estar necesariamente en este intervalo. Habíamos sugerido esta propiedad en la **Sección 2.3**.

3.3.2 La probabilidad de ocurrencia del espacio muestral es 1

Esto proviene del hecho de que el espacio muestral contiene todos los resultados posibles, por lo que al realizar un experimento siempre se debe obtener alguno de ellos.

Para ilustrar esta propiedad, consideremos nuevamente el experimento de extraer una ficha desde una urna, donde cada una de las fichas contenidas en ella muestra **1** de **4** símbolos posibles: círculo, triángulo, cruz o rombo. La **Figura V.13** muestra los símbolos observados en las fichas obtenidas en **6** extracciones realizadas desde la urna, donde la ficha extraída fue repuesta en la urna antes de hacer la siguiente extracción.

Figura V.13: Símbolos observados en las 6 fichas extraídas a partir de la urna, con reposición de cada ficha antes de la extracción de la siguiente.

Según la manera en que definimos la ocurrencia de un suceso, el espacio muestral ocurre si se obtiene un círculo, un triángulo, una cruz o un rombo, dado que todos estos resultados están en él. La **Tabla V.2** muestra, a través de tickets, la ocurrencia del espacio muestral a través de las **6** repeticiones del experimento, indicando que este siempre ocurrió. De acuerdo a esta tabla, la **Tabla V.3**, muestra la secuencia de frecuencias relativas a lo largo de las **6** repeticiones, todas iguales a **1**. Esta situación no cambiará, aunque aumentemos indefinidamente el número de repeticiones, por lo que la probabilidad que buscamos, o el valor al que se acercan cada vez más estas frecuencias relativas, debe ser igual a **1**, como habíamos anticipado.

Espacio muestral	Número de extracciones					
Resultado	1	2	3	4	5	6
○ △ + ◇	✓	✓	✓	✓	✓	✓

Tabla V.2: Ocurrencia del espacio muestral, a través de las 6 extracciones de una ficha con un símbolo, con reposición de la ficha después de cada extracción.

Espacio muestral	Número de extracciones					
Resultado	1	2	3	4	5	6
○ △ + ◇	$\frac{1}{1}=1$	$\frac{2}{2}=1$	$\frac{3}{3}=1$	$\frac{4}{4}=1$	$\frac{5}{5}=1$	$\frac{6}{6}=1$

Tabla V.3: Secuencia de frecuencias relativas de ocurrencia del espacio muestral, a través de las 6 extracciones de una ficha con un símbolo, con reposición de la ficha después de cada extracción.

3.3.3 Probabilidad de ocurrencia de 1 de 2 sucesos que no pueden ocurrir de manera simultánea

Para establecer esta propiedad, consideraremos nuevamente los resultados de las 6 extracciones de una ficha desde una urna, que se muestran en la **Figura V.13**. Las **Tablas V.4** y **V.5** organizan la información de modo de obtener las 4 secuencias de frecuencias relativas: de círculo, triángulo, cruz y rombo, a través de las 6 extracciones. Estas 4 secuencias se muestran en cada una de las filas asociadas a los símbolos de la **Tabla V.5**. Para su obtención, a modo de ejemplo, la **Tabla V.4** muestra que, luego de las 4 primeras extracciones, se ha obtenido 1 círculo, 1 triángulo, 2 cruces y ningún rombo. Las frecuencias relativas, hasta este momento, corresponden a $\frac{1}{4}$, $\frac{1}{4}$, $\frac{2}{4}$ y 0, para círculo, triángulo, cruz y rombo, respectivamente, como se muestra en la columna correspondiente a la cuarta extracción en la **Tabla V.5**, que se ha destacado en gris.

	Número de extracciones					
Resultado	1	2	3	4	5	6
○				✓		
△			✓		✓	
+	✓	✓				
◇						✓

Tabla V.4: Resultados observados en las 6 primeras extracciones de fichas marcadas en una urna, con reposición.

	Número de extracciones					
Resultado	1	2	3	4	5	6
○	0,00	0,00	0,00	0,25	0,20	0,17
△	0,00	0,00	0,33	0,25	0,40	0,33
+	1,00	1,00	0,67	0,50	0,40	0,33
◇	0,00	0,00	0,00	0,00	0,00	0,17
	1,00	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**

Tabla V.5: Secuencias de frecuencias relativas observadas para cada uno de los cuatro símbolos, en las 6 primeras extracciones, con reposición.

Consideremos ahora los sucesos "se obtiene un triángulo" y "se obtiene una cruz". Notemos que estos sucesos no pueden ocurrir simultáneamente en una misma realización del experimento, puesto que, si al extraer una ficha de la urna esta muestra un triángulo, entonces no puede salir una cruz, y lo mismo ocurre en el sentido inverso.

Estamos interesados en la probabilidad de que ocurra uno u otro de estos sucesos, es decir, en la probabilidad de que ocurra un triángulo o una cruz, por lo que en las **Tablas V.6** y **V.7** triángulo y cruz han sido integrados en un solo caso. La **Tabla V.6** destaca en gris la ocurrencia de triángulo o cruz, mientras que la **Tabla V.7** destaca en gris la secuencia de frecuencias relativas asociadas a la **Tabla V.6**. A modo de ejemplo, la **Tabla V.6** muestra que hasta la cuarta extracción, 3 veces ha ocurrido o un triángulo o una cruz, por lo que la frecuencia relativa de veces en que ha ocurrido uno u otro resultado corresponde a $\frac{3}{4}$ o **0,75**; valor que se indica en la **Tabla V.7**, en la columna asociada a la cuarta extracción.

	Número de extracciones					
Resultado	1	2	3	4	5	6
○				✓		
△ +	✓	✓	✓		✓	
◇						✓

Tabla V.6: Símbolos observados en las 6 primeras extracciones de fichas marcadas en una urna, con reposición. La segunda fila considera la ocurrencia de un triángulo o una cruz como un mismo caso.

	Número de extracciones					
Resultado	1	2	3	4	5	6
○	0,00	0,00	0,00	0,25	0,20	0,17
△ +	1,00	1,00	1,00	0,75	0,80	0,67
◇	0,00	0,00	0,00	0,00	0,00	0,17
	1,00	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**

Tabla V.7: Frecuencias relativas observadas en las 6 primeras extracciones, con reposición. La segunda fila considera la ocurrencia de un triángulo o una cruz como un mismo caso.

Mostraremos que existe una relación entre la frecuencia relativa de ocurrencia de triángulo o cruz, en la **Tabla V.7**, y las frecuencias relativas de ocurrencia de triángulo y de cruz, por separado, en la **Tabla V.5**. La **Figura V.14** muestra esta relación.

Resultado	Número de extracciones					
	1	2	3	4	5	6
○	0,00	0,00	0,00	0,25	0,20	0,17
△	0,00	0,00	0,33	0,25	0,40	0,33
+	1,00	1,00	0,67	0,50	0,40	0,33
◇	0,00	0,00	0,00	0,00	0,00	0,17
	1,00	1,00	1,00	1,00	1,00	1,00

△	0,00	0,00	0,33	0,25	0,40	0,33
+	1,00	1,00	0,67	0,50	0,40	0,33
Suma	1,00	1,00	1,00	0,75	0,80	0,67

Resultado	Número de extracciones					
	1	2	3	4	5	6
○	0,00	0,00	0,00	0,25	0,20	0,17
△ +	1,00	1,00	1,00	0,75	0,80	0,67
◇	0,00	0,00	0,00	0,00	0,00	0,17
	1,00	1,00	1,00	1,00	1,00	1,00

Figura V.14: Relación entre la suma de las frecuencias relativas de ocurrencia de triángulo y cruz por separado, y la frecuencia relativa de la ocurrencia de un triángulo o una cruz.

La **Figura V.14** destaca, en la tabla superior derecha, las secuencias de frecuencias relativas de triángulo y cruz por separado, que se muestran en la **Tabla V.5**, donde bajo ambas filas, aparece la suma de sus valores al término de cada una de las extracciones. En la misma figura, la tabla inferior resalta la secuencia de frecuencias relativas de obtención de triángulo o cruz, al integrar ambos símbolos en un solo caso, que también se muestran en la **Tabla V.7**. Vemos que los valores de esta última secuencia corresponden a la suma de las frecuencias relativas de cada símbolo por separado. Esta relación se mantendrá, no importando cuántas extracciones con reposición se realicen, por lo que podemos inferir que las probabilidades asociadas cumplirán con ella.

La relación que hemos encontrado se cumple cada vez que consideremos **2** sucesos que no pueden ocurrir de manera simultánea. En el siguiente apartado, veremos que esta relación no se cumple en otros casos.

En resumen

Según la definición frecuentista de probabilidad:

- Una probabilidad corresponde a un valor entre **0** y **1**.
- La probabilidad del espacio muestral es **1**.
- La probabilidad de que ocurra **1** de **2** sucesos que no pueden ocurrir simultáneamente corresponde a la suma de las probabilidades de ocurrencia de los **2** sucesos por separado.

Ejercicios

1. Andrea llevaba el registro de las frecuencias relativas de ocurrencia de cada resultado de 2 experimentos aleatorios, con el objeto de acercarse a los valores de las probabilidades de cada uno de ellos. Sin embargo, en un instante, Andrea confunde los papeles donde había anotado los últimos valores de las frecuencias relativas y ya no está segura de a qué experimento corresponde cada una de las frecuencias relativas anotadas en ellos. Los papeles indican los valores:

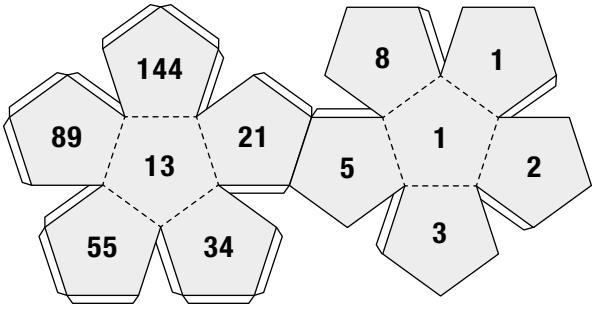
   0,3 0,375 0,225 0,375 0,375 0,01 0,025 0,29 0,025

   Si uno de los experimentos tiene 4 resultados posibles y el otro 5:

   a. ¿Cuáles podrían ser las frecuencias relativas asociadas a los resultados de cada uno de los 2 experimentos?
   b. En el apartado anterior, ¿puede verificar si existe una única respuesta?

2. En cada uno de los siguientes experimentos aleatorios, describa sucesos que no pueden ocurrir de manera simultánea:

   a. Se lanza 2 veces un dado equilibrado de 12 caras, numerado de la siguiente forma:



      i. Se registra la suma de los números de las caras superiores.
      ii. Se registra el producto de los números de las caras superiores.
      iii. Se registra la diferencia absoluta entre los números en las caras superiores.

   b. Se observa el tráfico de una calle y se registra la cantidad de autos que pasan por ella durante una hora.
   c. Se lanza un avión de papel y se registra la altura máxima que alcanza.

3.4. Probabilidad de ocurrencia de un suceso, otro o ambos, cuando es posible que ocurran los 2 de manera simultánea

En este apartado, estudiaremos la situación en la que los 2 sucesos de interés pueden ocurrir de manera simultánea. A modo de ejemplo, consideremos los sucesos "se obtiene un polígono" y "se obtiene un símbolo formado por entre 1 y 3 segmentos". Habiamos visto que, para que ocurra el primer suceso, el resultado del experimento debe ser triángulo o rombo, mientras que para que ocurra el segundo, el resultado del experimento debe ser triángulo o cruz. Vemos entonces que, si al extraer una ficha de la urna esta muestra un triángulo, entonces habrán ocurrido los 2 sucesos, dado que, por una parte, el símbolo corresponde a un polígono y, por otra, está formado por entre 1 y 3 segmentos.

Estudiaremos la relación entre las frecuencias relativas de los 2 sucesos de interés por separado, y la frecuencia relativa de ocurrencia de uno, otro o ambos. Para ello, debemos tener presente que:

- Ocurre el primer suceso si el símbolo obtenido es un triángulo o un rombo.
- Ocurre el segundo suceso si el símbolo obtenido es un triángulo o una cruz.
- Ocurre el primer suceso, el segundo o ambos, si el símbolo obtenido es un triángulo, una cruz o un rombo.

La **Figura V.15** muestra las frecuencias relativas de estos 3 sucesos a través de las 6 extracciones. Las tablas que se muestran han sido construidas en base a la información de la **Tabla V.4**. La **Figura V.15** destaca, en la tabla superior derecha, las secuencias de frecuencias relativas de los 2 sucesos de interés, por separado, mostrando bajo ambas filas la suma de sus valores al término de cada una de las extracciones. En la misma figura, la tabla inferior destaca la secuencia de frecuencias relativas de ocurrencia de uno, el otro o ambos sucesos. Vemos que la suma de las frecuencias relativas individuales ya no es siempre igual a la frecuencia relativa de ocurrencia de un suceso, el otro o los 2 a la vez, si no que, en algunos casos, es mayor que la última.

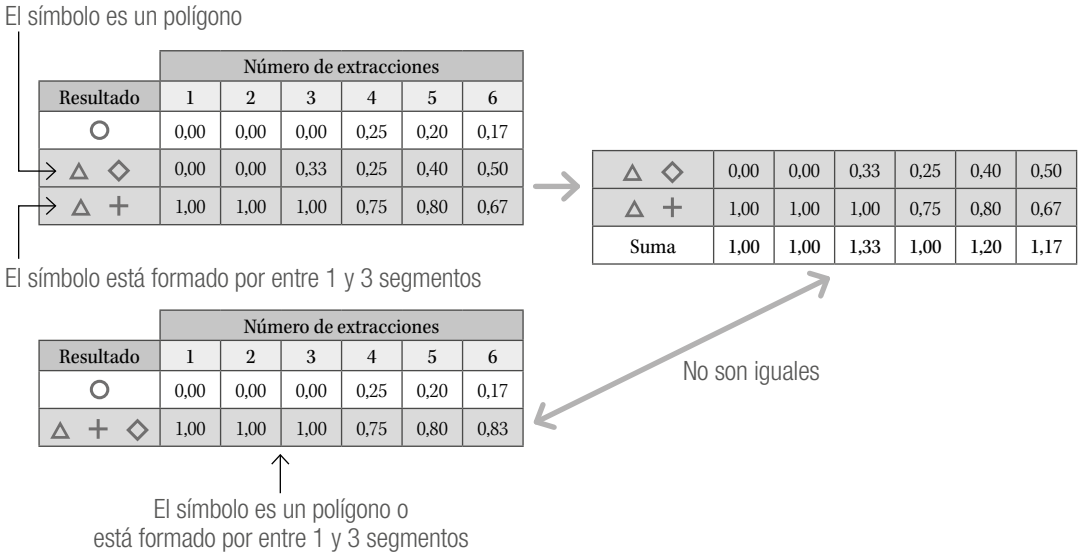


Figura V.15: Relación entre la suma de las frecuencias relativas de ocurrencia de los sucesos: "el símbolo en la ficha extraída es un polígono" y "el símbolo en la ficha extraída está formado por entre 1 y 3 segmentos", por separado, y la frecuencia relativa de ocurrencia de un suceso, el otro o ambos.

La diferencia entre las 2 situaciones ilustradas corresponde a que en el caso considerado en la **Sección 3.3.3**, los sucesos "se obtiene un triángulo" y "se obtiene una cruz", no contienen resultados en común, mientras que, en el segundo caso, los sucesos considerados, "el símbolo en la ficha extraída es un polígono" y "el símbolo en la ficha extraída está formado por entre 1 y 3 segmentos" tienen un elemento en común: el triángulo. De este modo, obtener un triángulo indica que ocurrió tanto el suceso "el símbolo en la ficha extraída es un polígono" como el suceso "el símbolo en la ficha extraída está formado por entre 1 y 3 segmentos". Esto hace que la suma de las probabilidades respectivas sea mayor que la probabilidad que nos interesa, que ocurra uno, otro o ambos sucesos, pues se está considerando 2 veces uno de los resultados, el triángulo.

Para corregir, entonces, debemos restar una vez la probabilidad de obtener un triángulo. A modo de ejemplo, notemos lo que ocurre al finalizar las 6 extracciones de fichas de la urna. En la **Tabla V.5** vemos, por ejemplo, que en la sexta extracción, la frecuencia relativa de triángulo era 0,33. Si restamos esta cantidad a la suma de las frecuencias relativas de los 2 sucesos de interés, obtenemos:

$$0{,}50 + 0{,}67 - 0{,}33 = 0{,}84$$

Ese valor coincide con la frecuencia relativa de la ocurrencia de un suceso, el otro o ambos, que se muestra en la **Figura V.15**[2].

Luego, esperaríamos que esta propiedad se transmitiera a las probabilidades respectivas, es decir, que para obtener la probabilidad de ocurrencia de 1 de 2 sucesos, o de ambos simultáneamente, se deben sumar las probabilidades de ocurrencia de cada uno de ellos por separado y restar a esta, la probabilidad de ocurrencia de ambos simultáneamente.

3.5. Probabilidad del suceso complemento

Esta propiedad se desprende de 2 resultados que ya vimos anteriormente:

- La probabilidad del espacio muestral debe ser 1.
- Dados 2 sucesos que no pueden ocurrir de manera simultánea, la probabilidad de que ocurra 1 de ellos corresponde a la suma de las probabilidades de ocurrencia de 1 de los 2 por separado.

Para ilustrar, consideremos, ahora, en el mismo experimento que seguimos, el suceso "el símbolo en la ficha extraída es un triángulo". Nos interesa estudiar la probabilidad de que este suceso no ocurra, es decir, de que "el símbolo en la ficha extraída no sea un triángulo". Para ello, notemos que el suceso no ocurre, si el símbolo en la ficha extraída es un círculo, una cruz o un rombo. La **Figura V.16** muestra la ocurrencia y no ocurrencia del suceso "el símbolo en la ficha extraída es un triángulo", a través de las 6 extracciones que seguimos.

[2] La Figura V.15 muestra el valor 0,83. La diferencia corresponde al redondeo a la centésima utilizado.

		Número de extracciones					
	Resultado	1	2	3	4	5	6
El suceso ocurre:	△			✓		✓	
El suceso no ocurre:	○ + ◇	✓	✓		✓		✓

		Número de extracciones					
	Resultado	1	2	3	4	5	6
El suceso ocurre:	△	0,00	0,00	0,33	0,25	0,40	0,33
El suceso no ocurre:	○ + ◇	1,00	1,00	0,67	0,75	0,60	0,67
	Suma	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**	**1,00**

Figura V.16: La suma de las frecuencias relativas de ocurrencia y no ocurrencia del suceso "se obtiene un triángulo" es 1.

Notamos que, después de cada una de las extracciones, la suma de las frecuencias relativas de ocurrencia y no ocurrencia del suceso "se obtiene un triángulo" es **1**. Al igual como hemos dicho con propiedades anteriores, esta propiedad se seguirá cumpliendo en la medida que continuemos realizando repeticiones del experimento. De esta forma, la propiedad descrita se transmitirá a las probabilidades respectivas.

Dado un suceso, cuando este no ocurre decimos que ha ocurrido su *suceso complemento*. De esta manera, el suceso complemento de un suceso dado corresponde al grupo de resultados del espacio muestral que no están en el suceso de interés. A modo de ejemplo, el suceso complemento del suceso "se obtiene un triángulo" corresponde a "se obtiene un círculo, una cruz o un rombo".

Luego, hemos encontrado que la suma de las probabilidades de ocurrencia de un suceso y de su complemento es siempre igual a **1**, lo que entrega una manera de obtener sus probabilidades. Por ejemplo, para obtener la probabilidad del suceso "se obtiene un círculo, una cruz o un rombo", podemos restar a **1** la probabilidad de obtener un triángulo.

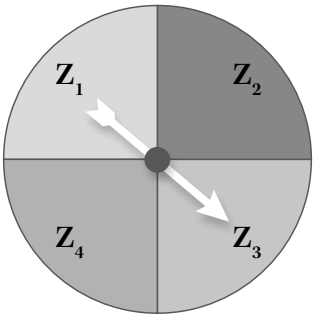
En resumen

Según la definición frecuentista de probabilidad, en un experimento aleatorio:

- Para obtener la probabilidad de que ocurra un suceso, otro o ambos simultáneamente, se deben sumar las probabilidades de ocurrencia de cada uno de ellos por separado, y restar la probabilidad de que ocurran los **2** de manera simultánea.
- La probabilidad de que ocurra un suceso puede obtenerse restando a **1** la probabilidad de que ocurra su complemento.

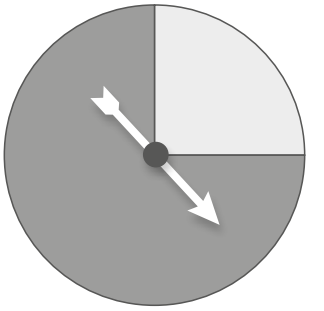
Ejercicios

1. Para la siguiente ruleta, suponga el experimento de girar la flecha y observar la región en que se detiene su extremo, Z_1, Z_2, Z_3 o Z_4. Para cada una de las afirmaciones que se presentan, determine si es verdadera o falsa. Justifique.



   a. La suma de las probabilidades de que la flecha se detenga en cada una de las zonas Z_1 a Z_4 es **1**.
   b. La probabilidad de que la flecha quede en una zona Z_5 es **0**.
   c. Para cada zona, la probabilidad de que la flecha se detenga en ella es menor que **1**.
   d. Las probabilidades de que la flecha quede en cada una de las zonas Z_1 a Z_4 son siempre mayores que **0**.

2. Pedro construyó una ruleta como la que se muestra en la figura, y consideró el experimento combinado de lanzar una vez una moneda y girar una vez la ruleta, registrando el lado que muestra la moneda al caer y el color en que se detiene la flecha de la ruleta.



   Después de repetir muchas veces el experimento, notó que al finalizar una realización de las **2** etapas, lanzar la moneda y girar una vez la ruleta, la suma de las frecuencias relativas de cara y de color oscuro era mayor a **1**. De acuerdo a lo descrito, responda:

   a. ¿Puede ser correcto lo que observó Pedro, o necesariamente debe haber un error en sus cálculos? Explique.
   b. ¿Cuál cree usted que es el valor de la probabilidad del suceso que describe Pedro?

3.6. Axiomas de probabilidad

Notemos que, a través del texto, hemos trabajado únicamente con la definición frecuentista de probabilidad. Si bien existen otras definiciones de este concepto, todas ellas persiguen el propósito de medir cuantitativamente la incerteza y, para ello, se ha establecido que, cualquiera sea la definición que se utilice, toda asignación de probabilidades a sucesos asociados a un experimento debe cumplir ciertas condiciones. Estas condiciones son llamadas los *3 axiomas de la probabilidad* y, aunque existen diferentes versiones, todas ellas son equivalentes y corresponden a las propiedades que hemos deducido, según la definición frecuentista en la sección anterior.

De este modo, los 3 axiomas de probabilidad pueden ser expresados, de manera coloquial, como:

- Una probabilidad debe ser un valor entre 0 y 1.
- La probabilidad del espacio muestral es 1.
- Si 2 sucesos no tienen resultados en común, es decir, no pueden ocurrir ambos a la vez, la probabilidad de ocurrencia de uno u otro suceso corresponde a la suma de las probabilidades de ocurrencia de cada uno de ellos por separado.

En resumen

Para toda definición de probabilidad, se debe cumplir que:

1. Una probabilidad corresponde a un valor entre 0 y 1.
2. La probabilidad del espacio muestral es 1.
3. La probabilidad de que ocurra un suceso u otro, cuando ellos no pueden ocurrir de manera simultánea, corresponde a la suma de las probabilidades de que ocurra cada suceso de manera individual.

En las Secciones 3.3 a 3.5 derivamos propiedades de las probabilidades al ser definidas de manera frecuentista. Las 3 primeras de ellas corresponden exactamente a los 3 axiomas de la probabilidad. En las secciones referidas, mostramos también que la definición frecuentista de probabilidad indica una manera de obtener probabilidades de ocurrencia de un suceso, otro o ambos, cuando ellos pueden ocurrir de manera simultánea, y de obtener la probabilidad de no ocurrencia de un suceso, o de ocurrencia de su complemento. Veremos aquí que estos resultados son válidos bajo cualquier definición de probabilidad, una vez que se cumplen los 3 axiomas que hemos establecido.

Consideremos, a modo de ejemplo, el lanzamiento de un dado y los sucesos "se obtiene un número par" y "se obtiene un múltiplo de 3". Claramente, los sucesos pueden ocurrir de manera simultánea, puesto que, en caso de obtenerse un 6, ambos habrán ocurrido. Luego, para obtener la probabilidad de ocurrencia de 1, el otro o ambos, no es posible utilizar el tercer axioma de la probabilidad.

Notemos que el primer suceso, "se obtiene un número par", ocurre, si el número obtenido es 2, 4 o 6. Luego, por el tercer axioma de la probabilidad, dado que estos 3 resultados no pueden ocurrir de maneara simultánea, la probabilidad de que se obtenga un número par corresponde a la suma de las probabilidades de que se obtenga un 2, un 4 y un 6, lo que anotaremos como:

$$p_2 + p_4 + p_6$$

Del mismo modo, el segundo suceso, "se obtiene un múltiplo de 3", ocurre, si el número obtenido es 3 o 6, por lo que la probabilidad del segundo suceso es:

$$p_3 + p_6$$

Por otra parte, ocurre al menos 1 de los 2 sucesos cuando el número observado en la cara superior es 2, 3, 4 o 6, por lo que la probabilidad buscada es:

$$p_2 + p_3 + p_4 + p_6$$

Notamos que, al sumar las probabilidades de los 2 sucesos por separado, habremos sumando 2 veces la probabilidad de obtener un 6, p_6, por lo que será necesario restarla para obtener la probabilidad de interés:

$$(p_2 + p_4 + p_6) + (p_3 + p_6) - p_6 = p_2 + p_3 + p_4 + p_6$$

Es decir, debemos sumar las probabilidades de ocurrencia de cada suceso por separado y restar la probabilidad de ocurrencia de ambos a la vez.

En resumen

Para toda definición de probabilidad:

- La probabilidad de que ocurra un suceso, otro o ambos simultáneamente, corresponde a la suma de las probabilidades de ocurrencia de cada uno de ellos por separado, a la que debe restarse la probabilidad de que ocurran los 2 sucesos de manera simultánea.

Para ver qué ocurre con la probabilidad del complemento del suceso, consideremos el suceso "se obtiene un múltiplo de 3". Por la argumentación realizada anteriormente, la probabilidad de este suceso corresponde a:

$$p_3 + p_6$$

Por otra parte, el suceso de interés no ocurre, u ocurre su complemento, si el número observado corresponde a 1, 2, 4 o 5. Luego, la probabilidad de que el suceso no ocurra corresponde a:

$$p_1 + p_2 + p_4 + p_5$$

Con el mismo razonamiento, sabemos que la probabilidad del espacio muestral completo corresponde a:

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6$$

Ya sabemos, por el segundo axioma, que esta suma debe ser igual a 1. Luego, la suma de las probabilidades de ocurrencia del suceso de interés, $p_3 + p_6$, y de su complemento, $p_1 + p_2 + p_4 + p_5$, es igual a 1.

En general, este resultado nos muestra una manera de obtener la probabilidad de ocurrencia de un suceso restando a 1 la probabilidad de ocurrencia de su complemento.

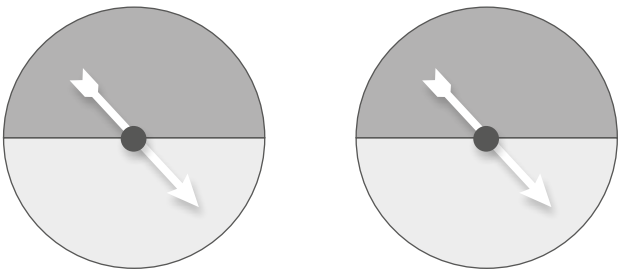
En resumen

Para toda definición de probabilidad:

- La probabilidad de ocurrencia de un suceso puede obtenerse restando a 1 la probabilidad de ocurrencia de su complemento.

Ejercicios

1. Considere las ruletas en la figura:



Elisa gana si, al girar las ruletas, ambas flechas indican las zonas de color más claro. Ella piensa que tiene un 50% de probabilidades de ganar. ¿Está usted de acuerdo con Elisa? Justifique su respuesta.

2. Se lanza una moneda equilibrada 3 veces.
   a. ¿Cuál es la probabilidad de que se obtengan 2 caras y 1 sello, sin importar el orden en que se obtuvieron?
   b. ¿Cuál es la probabilidad de que se obtenga cara en primer y segundo lugar, y sello en tercer lugar?
   c. ¿A qué cree usted que se debe que las probabilidades en a. y b. difieran?

3.7. Equiprobabilidad, conteo y uso de árboles

Consideremos el experimento de lanzar un dado y registrar el número observado en su cara superior, y sus resultados posibles, 1, 2, 3, 4, 5 o 6. Podemos convenir que todos estos resultados son igualmente probables, pues siendo el dado equilibrado, no hay motivos para pensar que uno de los números va a ocurrir más frecuentemente que los otros. En este caso, decimos que *los resultados posibles son equiprobables*, o que *el espacio muestral es equiprobable*.

Consideremos cada uno de los resultados, 1, 2, 3, 4, 5 y 6, por separado y notemos que:

- Ya que ningún par de resultados puede ocurrir de manera simultánea, la probabilidad de ocurrencia de uno u otro de ellos corresponde a la suma de sus probabilidades de ocurrencia individuales.
- La ocurrencia de uno u otro es equivalente a la ocurrencia del espacio muestral completo. Por lo tanto, *la probabilidad de ocurrencia de uno u otro de ellos* es igual a 1.

Si agregamos ahora el supuesto de equiprobabilidad, es decir, que las probabilidades de cada uno de los 6 resultados son iguales, las afirmaciones anteriores implican que ellas son todas iguales a $\frac{1}{6}$.

En efecto, si anotamos como p la probabilidad de obtener cada uno de los resultados 1, 2, 3, 4, 5 y 6, la primera afirmación dice que la probabilidad de ocurrencia de uno u otro de ellos es:

$$p + p + p + p + p + p = 6p$$

La segunda afirmación dice que esta misma probabilidad debe ser igual a 1. Luego, se debe cumplir que:

$$6p = 1$$

Es decir, $p = \frac{1}{6}$. En el caso general, podemos decir que si los resultados del experimento son equiprobables, la probabilidad de ocurrencia de cada uno de ellos corresponde a:

$$\frac{1}{\textit{número de resultados posibles}}$$

De este modo, si en un curso hay 30 niños y escogemos un niño de manera aleatoria, la probabilidad de que un niño en particular digamos, Pedro, sea el elegido corresponde a $\frac{1}{30}$.

Por otra parte, si en una bolsa tenemos 10 bolitas numeradas del 1 al 10 y extraemos una de ellas, existen 10 resultados posibles. Si todas las bolitas son del mismo tamaño y textura, y no es posible mirar dentro de la bolsa antes de la extracción, podemos convenir que los 10 resultados son equiprobables. En este caso, la probabilidad de que se obtenga, por ejemplo, el número 8, es $\frac{1}{10} = 0{,}1$.

Supongamos que nos interesa conocer, por ejemplo, la probabilidad de que el número de la bolita extraída sea par, es decir, nos interesa la probabilidad de que salga 2, 4, 6, 8 o 10. Cada uno de estos números por separado tiene probabilidad $\frac{1}{10}$ de ocurrir, como discutimos anteriormente. Dado que una pareja de ellos no puede ocurrir al mismo tiempo, la probabilidad de que salga 2, 4, 6, 8 o 10 es la suma de las probabilidades de cada uno de estos 5 resultados por separado, es decir, corresponde a:

$$\frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{5}{10} = 0{,}5$$

Notamos que el resultado corresponde a la suma de la probabilidad que tiene cada resultado, en este caso, $\frac{1}{10}$, tantas veces como resultados posea el suceso de interés, en este caso, 5. De este modo, en el numerador obtenemos el número de resultados que son favorables al suceso, mientras que en el denominador obtenemos el número de resultados en el espacio muestral completo. Podemos resumir este resultado diciendo que si los resultados de un espacio muestral son equiprobables, la probabilidad de ocurrencia de un suceso cualquiera asociado a dicho espacio muestral corresponde a:

$$\frac{\textit{número de resultados favorables al suceso}}{\textit{número de resultados posibles}}$$

Para reforzar esta idea, planteamos un nuevo ejemplo. Consideremos el lanzamiento de 3 monedas de manera consecutiva. Como hemos visto, existen 8 resultados posibles:

CCC CCS CSC CSS SCC SCS SSC SSS

El suceso "se obtienen exactamente 2 caras" ocurre si el resultado del experimento es 1 de los 3 resultados

CSC CCS SCC

Luego, la probabilidad de que se obtengan exactamente 2 caras corresponde a

$$\frac{3}{8} = 0{,}375$$

Notemos que, con lo anterior, hemos dado una fórmula de cálculo de probabilidades cuando todos los resultados del experimento tienen la misma probabilidad de ocurrir. En la práctica, si esto es o no así en un experimento dado, debe ser juzgado con criterio y según nuestra experiencia y, en general, ello refleja la simetría de la obtención de los resultados.

A modo de ejemplo, en general, asumimos que en un dado de 6 caras, cada lado tiene la misma probabilidad de quedar en la cara superior después de un lanzamiento. Sin embargo, existen dados no equilibrados en los que, por ejemplo, algunas caras tienen mayor superficie que otras. En estos casos, no todas las caras tienen la misma probabilidad de quedar en la cara superior al lanzar el dado. Nuestra asignación de probabilidades no debiese ser la misma para todas las caras y no podremos calcular probabilidades en base al modelo "resultados favorables dividido por resultados posibles".

En resumen

Cuando los resultados de un experimento son *equiprobables*, es decir, tienen la misma probabilidad de ocurrir, podemos obtener la probabilidad de un suceso como

$$\frac{\textit{número de resultados favorables al suceso}}{\textit{número de resultados posibles}}$$

A este resultado se le conoce como *regla de Laplace*.

El resultado que encontramos requiere de contar el número de resultados en un suceso. Cuando este número es demasiado grande, deja de ser razonable hacerlo a través del listado de todos los resultados, como lo hemos hecho hasta ahora, y es necesario conocer ciertas reglas para obtener estas cantidades.

Una de ellas corresponde al llamado *principio multiplicativo*. Para derivarlo, supongamos el ejercicio de lanzar un dado y una moneda y registrar el número resultante en el dado y el lado obtenido en la cara superior de la moneda. ¿Cuántos son los resultados posibles de este experimento? Para derivar el principio multiplicativo, comenzaremos identificando estos resultados.

Notemos, primero, que experimentos como el descrito pueden ser desglosados en etapas o subexperimentos. En el ejemplo, podemos desglosar el experimento en 2 etapas:

i. Lanzar el dado.
ii. Lanzar la moneda.

Este enfoque presenta ventajas, tanto para visualizar, como para contar los resultados posibles. En efecto, notemos que, en el ejemplo, la primera etapa del experimento, lanzar el dado, tiene 6 resultados posibles: los números enteros del 1 al 6. Independientemente de cuál haya sido el resultado en esta etapa, la segunda etapa, lanzar la moneda, tiene 2 resultados posibles: cara y sello. Esta situación se muestra en la **Figura V.17**.

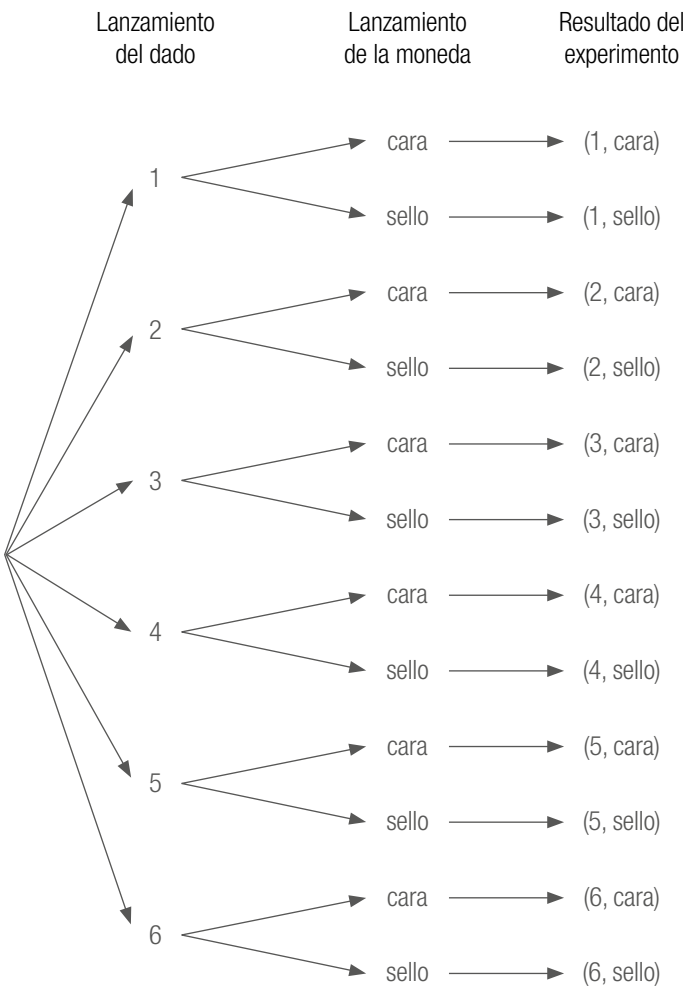


Figura V.17: Resultados posibles en el lanzamiento de un dado y una moneda.

La figura ayuda a visualizar la idea de que "para cada uno de los **6** resultados posibles en el lanzamiento del dado, existen **2** resultados posibles en el lanzamiento de la moneda". Esto indica que, para obtener el número total de resultados del experimento debemos, multiplicar el número de resultados en cada una de las dos etapas, **6 • 2 = 12**. En efecto, la columna a la derecha en la **Figura V.17** lista los resultados posibles del experimento, donde se verifica que estos son **12**. Esta estrategia para obtener el número de resultados de un experimento se conoce como el *principio multiplicativo*, dado que se multiplican los números de resultados posibles en cada una de las etapas.

La representación de los resultados de un experimento utilizada en la **FiguraV.17** se denomina *diagrama de árbol*.

Para fijar las ideas, consideremos otro ejemplo. Un día dado, el menú del casino ofrece 2 posibles platos de entrada, palta rellena o tomate relleno, 3 posibles platos de fondo, porotos granados, pollo o pescado, y 2 posibles postres, manzana o flan. Si para formar su almuerzo, un niño debe elegir exactamente 1 entrada, 1 plato de fondo y 1 postre, ¿cuántos son los menús posibles? La **Figura V.18** muestra esta situación.

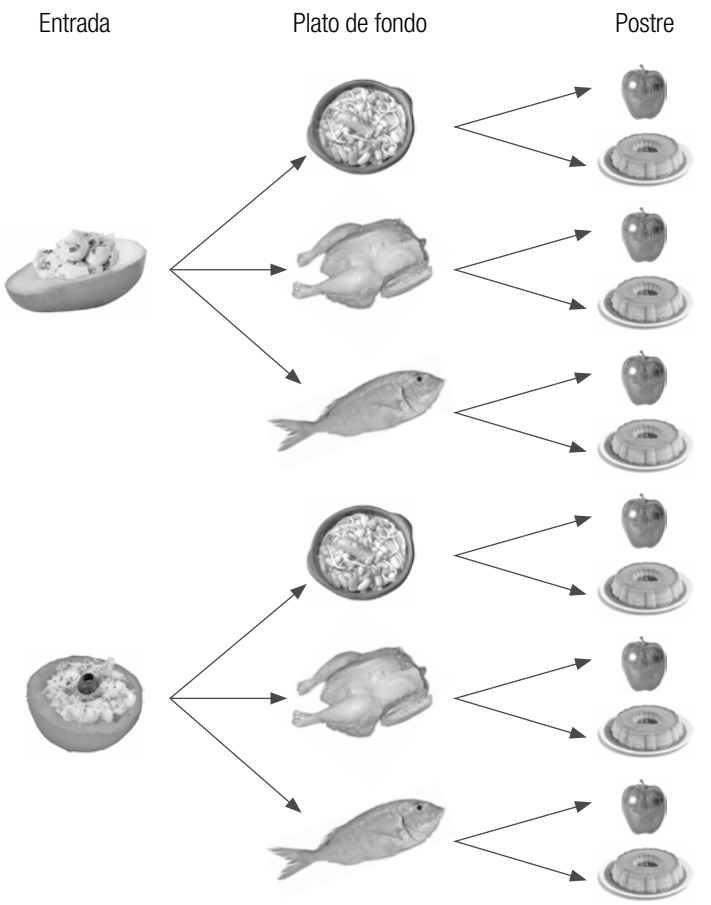


Figura V.18: Posibles menús creados con 2 alternativas de entrada, 3 alternativas de plato de fondo y 2 alternativas de postre.

Podemos pensar en el experimento de elegir un almuerzo, desglosándolo en 3 etapas, que corresponden a la elección de entrada, de plato de fondo y de postre. Luego, el principio multiplicativo dice que el número total de menús corresponde a "2 entradas" • "3 platos de fondo" • "2 postres" = 2 • 3 • 2 = 12 menús.

Notemos que, en casos como los anteriores, el orden en que se consideren las etapas es irrelevante para determinar el número de resultados posibles. En efecto, en el primer ejemplo, si en lugar de considerar primero el lanzamiento del dado consideramos primero el de la moneda, el principio multiplicativo dice que esto puede ocurrir de "2 lados de la moneda" • "6 lados del dado" = 2 • 6 = 12 formas, igual al número de formas determinado anteriormente. Lo mismo ocurre en la situación del casino: podríamos elegir primero el postre y luego la entrada y el plato de fondo, y el número de resultados posibles seguiría siendo 12. Esto se debe a que la multiplicación es conmutativa.

En otro ejemplo, consideremos 3 niños que desean tomar turnos para andar en una patineta. Para ello, deciden poner 3 papeletas en una urna: una con cada uno de sus nombres, Catalina, Daniela y Joaquín, y extraer las papeletas de manera aleatoria, de modo que utilizarán la patineta en el orden en que se obtengan los nombres al extraer las papeletas. El número de formas en que ellos tomen turnos, es decir, de ordenamientos en que pueden usar la patineta, corresponde al número de formas en que estas 3 papeletas pueden ser extraídas desde la urna, para lo que utilizaremos la regla multiplicativa.

La **Figura V.19** muestra esta situación en un diagrama de árbol. En la primera extracción, existen 3 resultados posibles: los nombres de cada uno de los 3 niños. Una vez elegida la primera papeleta, independientemente de cuál de ellas haya sido elegida, en la segunda extracción existen solo 2 resultados posibles, una con cada uno de los nombresde los 2 niños que aún no han sido seleccionados. Finalmente, en la tercera extracción, solo existe un resultado posible, independientemente de los resultados de las etapas anteriores.

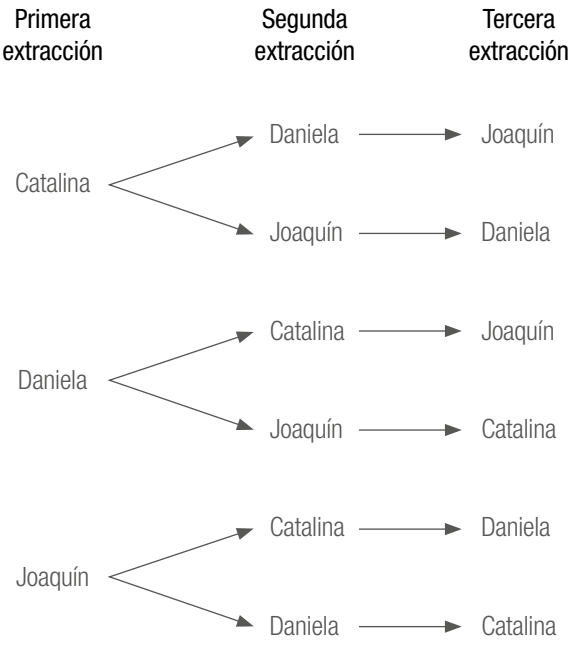


Figura V.19: Maneras de repartir los turnos para andar en patineta de 3 niños.

En la **Figura V.19** notamos que los resultados posibles en la segunda extracción no son los mismos, ya que dependen de cuál fue el resultado en la primera etapa. En efecto, si en la primera extracción se obtuvo el nombre de Catalina, este no es un resultado posible en la segunda extracción. Sin embargo, sí lo es cuando el nombre extraído en la primera etapa es Daniela o Joaquín. Hacemos notar que esto no invalida la utilización de la regla multiplicativa, dado que lo que en realidad interesa es el número de resultados posibles en cada una de las etapas, no cuáles son ellos.

De este modo, el número de ordenamientos posibles para utilizar la patineta son:

$$3 \cdot 2 \cdot 1 = 6$$

Si queremos, ahora, calcular la probabilidad de que Daniela ocupe el segundo turno, notamos que en la primera etapa son resultados favorables los nombres de Catalina y Joaquín, es decir, 2 resultados. En la segunda etapa, solo existe un resultado favorable, que corresponde al nombre de Daniela, mientras que en la tercera etapa, solo existe un resultado posible, que es el nombre de quien aun no ha sido elegido. De este modo, el número de casos favorables al evento de interés es:

$$2 \cdot 1 \cdot 1 = 2$$

Con esto, la probabilidad de que Daniela ocupe el segundo turno es:

$$\frac{\textit{número de resultados favorables al suceso}}{\textit{número de resultados posibles}} = \frac{2}{6} \approx 0{,}33$$

Notemos que un diagrama de árbol corresponde a un recurso que ayuda a comprender el principio multiplicativo pero que, al crecer el número de etapas, su uso se vuelve ineficiente. En efecto, si el número de niños que desea tomar turnos para utilizar la patineta fuese 10, en lugar de 3, un diagrama de árbol resultaría inmanejable. Sin embargo, por la regla multiplicativa, sabemos que el número de maneras de tomar turnos de estos 10 niños es

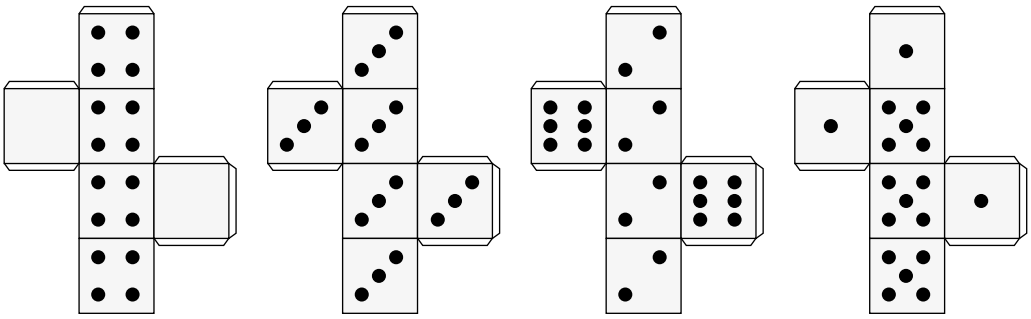
$$10 \cdot 9 \cdot 8 \cdot \ \ldots \ \cdot 3 \cdot 2 \cdot 1$$

En resumen

- Podemos utilizar una *representación de árbol* para determinar el espacio muestral asociado a un experimento.
- *Regla multiplicativa*: si un experimento puede ser desglosado en etapas, el número total de resultados del experimento completo corresponde a la multiplicación del número de resultados de cada una de las etapas.

Ejercicios

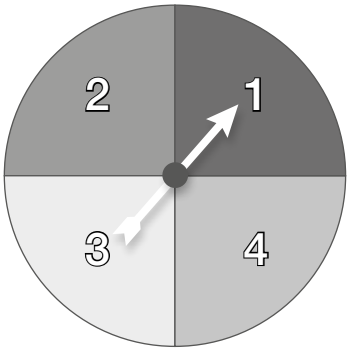
1. Utilizando las plantillas que se muestran, Andrea construye 4 dados equilibrados diferentes.



Para cada dado, Andrea piensa que sus resultados posibles no son equiprobables.

a. ¿Tiene razón Andrea? Justifique su respuesta.
b. Para cada dado, indique la probabilidad de cada uno de sus resultados posibles.

2. Se hace girar 2 veces una ruleta como la que se muestra en la figura:



Determine la probabilidad de que:

a. Ambos números obtenidos sean números pares.
b. La suma de los números obtenidos sea menor a 4.
c. El producto de los números obtenidos sea impar.

3. Suponga que se lanza un dado equilibrado de 6 caras. Obtenga las probabilidades de:

a. Obtener un número par.
b. Obtener un número primo o impar.
c. Obtener un número par o primo.
d. Obtener un cuadrado perfecto o par.
e. Obtener un número triangular o un número pentagonal.

4. Joaquín resuelve 5 de los 8 problemas que le ha solicitado su profesor de matemática. Al día siguiente, el profesor lo llama a la pizarra y le pide que muestre la solución de uno de los problemas asignados, que él elige extrayendo de manera aleatoria un número desde una caja que contiene papeles con los números del 1 al 8.
   a. ¿Cuál es la probabilidad de que Joaquín haya resuelto el problema elegido?
   b. ¿Son equiprobables los sucesos "Joaquín ha resuelto el problema elegido" y "Joaquín no ha resuelto el problema elegido"? Justifique.

5. Suponga que se extrae de manera aleatoria una bolita desde una urna que contiene 3 bolitas rojas, 5 azules y 2 blancas. Obtenga las probabilidades de:
   a. Obtener una bolita azul o blanca.
   b. Obtener una bolita que no sea roja.
   c. Obtener una bolita azul o verde.

6. Una ruleta está dividida en tres sectores con igual superficie, indicados con los números 1, 2 y 3. Suponga que se gira dos veces la ruleta. Encuentre la probabilidad de que:
   a. Se obtenga el número 3 al menos una vez.
   b. Se obtenga un número mayor a 5 al sumar los resultados de cada una de las vueltas.
   c. Se obtenga un número menor a 1 al sumar los resultados de cada una de las vueltas.

7. Un dado y una moneda son lanzados al mismo tiempo. Obtenga las probabilidades de que:
   a. Se obtenga una cara en la moneda.
   b. Se obtenga un número impar en el dado.
   c. Se obtenga una cara en la moneda y un número impar en el dado.
   d. Se obtenga un sello o un número primo.

8. Considere una caja A, que contiene 4 fichas rojas, 3 fichas verdes y 2 fichas azules, y otra caja B, que contiene 3 fichas rojas y 2 fichas verdes. Si se selecciona una ficha de manera aleatoria desde cada una de las cajas, determine la probabilidad de que:
   a. Una ficha sea roja y la otra sea azul.
   b. Una ficha sea roja y la otra sea verde.
   c. Las 2 fichas sean del mismo color.

9. Los 7 niños de un club desean formar una directiva compuesta por un presidente, un secretario y un tesorero. Ellos deciden hacerlo extrayendo 3 números a partir de una urna en la que ponen siete papeletas con los números del 1 al 7, donde cada número representa a uno de ellos. La primera papeleta elegida indicará quién será el presidente, la segunda indicará quién será el secretario y la tercera, quién será el tesorero. Determine:
   a. ¿De cuántas maneras distintas puede formarse la directiva?
   b. Si 4 de los niños son hombres y 3 son mujeres, ¿cuál es la probabilidad de que la directiva quede conformada solo por mujeres?
   c. ¿Cuál es la probabilidad de que una mujer sea elegida presidenta, y de que el secretario y el tesorero sean hombres?

10. Para subir una montaña existen 5 caminos.

    a. ¿De cuántas maneras puede un turista subir la montaña y luego bajarla, si puede hacerlo por 2 caminos diferentes, eligiendo cada uno de ellos de manera aleatoria entre los 5 caminos disponibles?

    b. ¿Cuál es la probabilidad de que el turista suba y baje por el mismo camino?

    c. ¿Cuál es la probabilidad de que suba y baje por caminos diferentes?

11. Determine cuántos números enteros entre 100 y 999 cumplen que todos sus dígitos son impares y distintos. Según esto, determine la probabilidad de que:

    a. El número tenga un número primo en las centenas.

    b. El número sea divisible por 3.

12. Un comité de la escuela está formado por un profesor, 3 alumnos y 4 alumnas. Si se seleccionan al azar 2 representantes de este comité para dar el discurso de bienvenida a los alumnos nuevos al comenzar el año, determine la probabilidad de que:

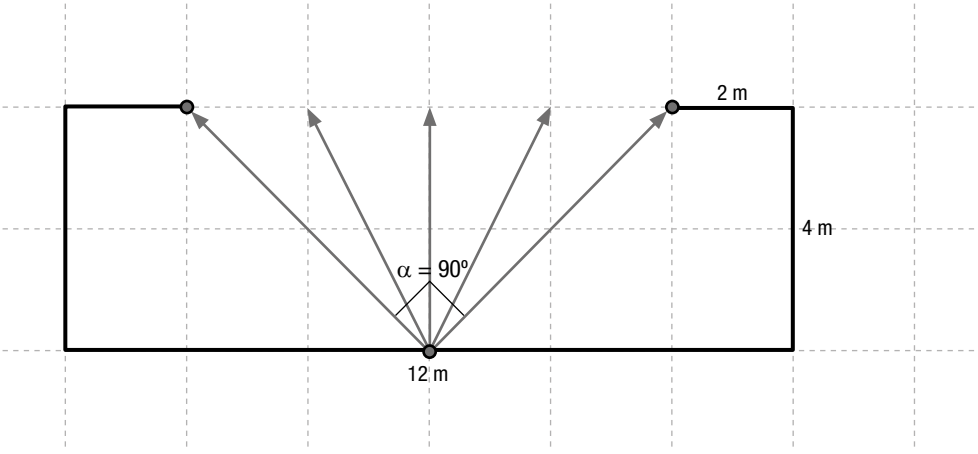
    a. Ambas sean alumnas.

    b. Ambas no sean alumnas.

    c. Uno de los representantes sea alumno y el otro sea alumna.

    d. Se elija al profesor y un alumno.

13. Un niño juega a lanzar una pelota desde el fondo de una habitación, como se muestra en la figura:



El niño elige de manera aleatoria la dirección a la que dirigirá la pelota. ¿Cuál es la probabilidad de que, al lanzar una vez la pelota, esta salga por la puerta de la habitación, asumiendo que no da botes en las paredes y que el lanzamiento es suficientemente fuerte?

Ayuda: considere obtener las medidas de los ángulos que se muestran.

Ejercicios del capítulo

1. Discuta si las siguientes situaciones son o no experimentos aleatorios para la persona que registra su resultado.

   a. Un niño multiplica 3 por 7 en una calculadora de bolsillo y registra su resultado.

   b. Sebastián corre la prueba de 100 metros planos y registra el tiempo que demora en llegar a la meta.

2. En cada una de las siguientes situaciones, proponga elementos que las complementen, permitiendo definir completamente un experimento aleatorio.

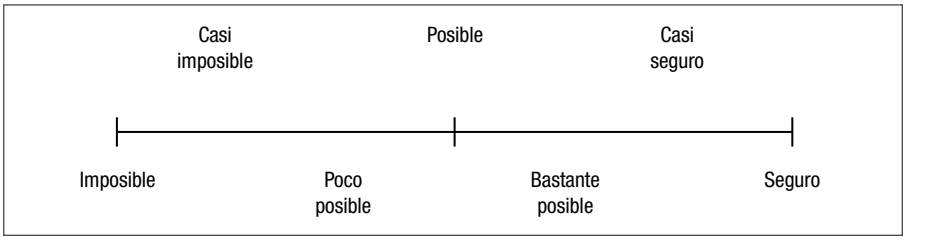
   a. María visita un jardín botánico.

   b. Un encuestador registra una respuesta entregada por un transeúnte.

   c. Carlos llama por teléfono a la casa de su abuela.

   d. Un niño juega con un dado.

3. Considere las siguientes situaciones e indique el grado de posibilidad que usted asignaría a su ocurrencia. Especifique los supuestos que realiza para evaluar dicho grado. Puede utilizar la escala:



   a. Que al contar los fósforos en una caja, usted encuentre exactamente 53.

   b. Que al dirigirse al cine, una persona se encuentre allí con uno o más conocidos.

   c. Que al conectarse a Facebook, usted encuentre conectados a 3 de sus amigos.

   d. Que al abrir un libro de 100 páginas, usted quede entre las páginas 50 y 70.

4. Considere 4 cartas numeradas con los números 2, 3, 5 y 7, respectivamente. Si se seleccionan 2 cartas diferentes de manera aleatoria. Obtenga la probabilidad que:

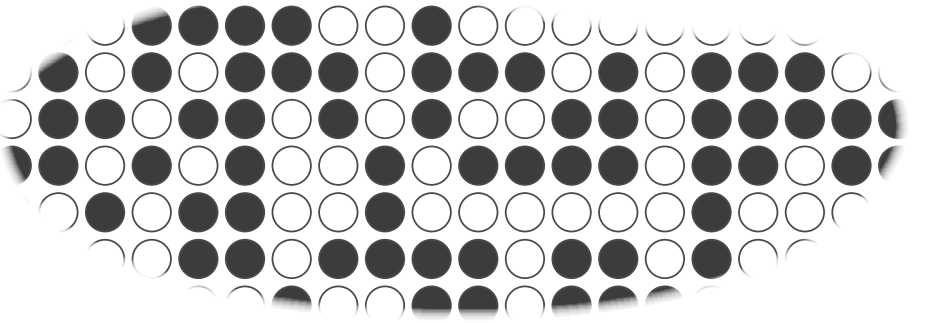
   a. La suma de los números de las cartas elegidas sea un número impar.

   b. El producto de los números de las cartas elegidas sea un número impar.

5. Pedro llevaba el registro del color de las bolitas que se extraían de un expendedor que contenía bolitas negras y blancas. Él registró la extracción de **130** bolitas, las cuales eran devueltas al expendedor después de cada extracción.

   a. Al finalizar la extracción de las **130** bolitas, Pedro obtuvo que la frecuencia relativa de bolitas de color negro era **0,60**. ¿Cuál era entonces la frecuencia relativa de bolitas de color blanco en las mismas **130** extracciones?

   b. Pedro llevaba el registro del color de la bolita extraída pintando un círculo de dicho color. Sin embargo, al finalizar **2.000** extracciones, notó que su representación se había hecho confusa, como en la figura:



   Aproximadamente, ¿cuántas bolitas de color negro y cuántas bolitas de color blanco cree usted que obtuvo Pedro, considerando la frecuencia relativa en a.?

   c. Dada la confusión, Pedro comenzó nuevamente las repeticiones del experimento, anotando las frecuencias relativas después de cada extracción en una tabla, como la siguiente:

Número de repeticiones	Frecuencia relativa de bolitas negras
50	0,680
100	0,620
200	0,675
300	0,660
400	0,658
500	0,648
600	0,658
700	0,661
800	0,670
900	0,653
1.000	0,651

   En base a los valores de la tabla, aventure las probabilidades de extraer una bolita blanca y de obtener una bolita negra en una extracción a partir del mismo expendedor.

6. Considere las siguientes actividades para un recreo:

a. Proponga un experimento aleatorio que los niños puedan realizar, basándose en estas actividades.

b. Para cada uno de los experimentos propuestos, describa un espacio muestral.

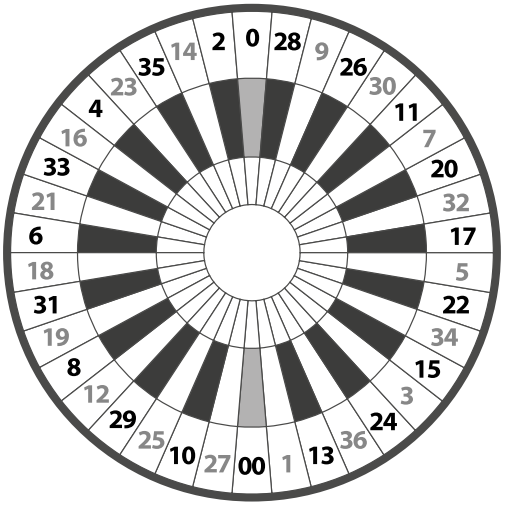
c. Describa 2 actividades que los niños puedan desarrollar dentro de la sala de clases, que puedan ser utilizadas para realizar un experimento aleatorio.

d. Describa un espacio muestral para cada uno de los experimentos propuestos.

7. En cada uno de los siguientes experimentos aleatorios, describa sucesos que no pueden ocurrir de manera simultánea:

a. Se observa un volantín en vuelo y se registra el número de vueltas que da en 1 minuto.

b. Se hace rodar una bolita en la ruleta de la figura, y se registran el color y el número en el cual cayó la bolita.



8. Resuelva los siguientes problemas justificando cada uno de sus pasos a partir de los 3 axiomas de probabilidad, y los resultados derivados de estos en la **Sección 3.6**.

   a. En una frutera hay 3 naranjas, 4 plátanos y 6 manzanas. Si se sacan 2 frutas al azar, ¿cuál es la probabilidad de que ambas sean manzanas, o ambas sean plátanos?

   b. Se extrae 2 bolitas desde una bolsa que contiene 5 bolitas marcadas con los números del 1 al 5.

      i. ¿Cuál es la probabilidad, de obtener un 2 y un 3, si la primera bolita extraída no es devuelta a la bolsa antes de extraer la segunda?

      ii. ¿Cuál es dicha probabilidad si la primera bolita es devuelta antes de extraer la segunda?

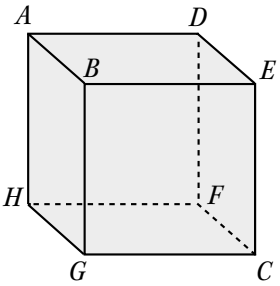
9. Considere 2 grupos de números, que denominaremos C y D. El grupo C está compuesto por los números 2, 3 y 6, mientras que el grupo D está compuesto por los números 3, 5 y 7. Un número es seleccionado al azar desde el grupo C y es denotado por "c". Otro número es seleccionado al azar desde grupo D, y es denotado como "d". Encuentre la probabilidad que la fracción $\frac{c}{d}$ sea:

   a. Menor que 0,5.

   b. Mayor que 1.

   c. Mayor que 0,5 y menor que 1.

10. Una hormiga muy particular se pasea por las aristas de un cubo. Al llegar a cada vértice, la hormiga cambia de arista y elige una arista al azar, sin volver atrás. La hormiga parte desde la arista indicada con la letra A.



   a. ¿Cuál es la probabilidad de que, tras recorrer 3 aristas, llegue a la arista indicada con la letra C?

   b. ¿Cuál es la probabilidad de que, tras recorrer 3 aristas, llegue a la arista indicada con la letra D?

BIBLIOGRAFÍA

Batanero, C. y Díaz, C. (Eds.). *Estadística con proyectos*. Departamento de Didáctica de la Matemática. Granada. España. 2011.

Batanero, C Burrill, G., y Reading, C. *Teaching statistics in school mathematics.-Challenges for teaching and teacher education. A Joint ICMI/IASE Study*. ICMI Study volume 14. Springer. New York. 2011.

Batanero, C. *Estadística y didáctica de la matemática: relaciones, problemas y aportaciones mutuas* en Penalva, C., Torregrosa, G. y Valls, J. (Eds.), *Aportaciones de la didáctica de la matemática a diferentes perfiles profesionales* (pp. 95-120). Universidad de Alicante. Granada. España. 2002.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck R., Perry, M., Scheaffer, R. *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum*. 2007. Disponible en: http://www.amstat.org/education/gaise/GAISEPreK12_Intro.pdf

Garfield, J. y Ben-Zvi, D. (Eds.). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*. Springer. USA. 2008.

Moore, D. *Statistics: Concepts and Controversies*. 6ª edición. W. H. Freeman and Company, New York. USA. 2009.

Ministerio de Educación. *Bases Curríulares de primero a sexto básico*. Recuperado el 1 de marzo del 2013. Disponible en: http://www.mineduc.cl/index5_int.php?id_portal=47&id_contenido=17116&id_seccion=3264&c=1

Minesterio de Educación. *Estándares Orientadores Para Egresados De Carreras De Pedagogía En Educación Básica*. Recuperado el 1 de marzo del 2013. Disponible en: http://www.mineduc.cl/usuarios/cpeip/File/2012/librobasicaokdos.pdf

Reys, R. *Helping Children Learn Mathematics*. 9ª edición. Wiley. USA. 2004.

Sowder, J., Sowder, L. y Nickerson S. *Reconceptualizing Mathematics for Elementary School Teachers*: Instructor's Edition. W. H. Freeman and Company, New York. USA. 2009.

Utts, J. y Heckard, R. *Mind on Statistics*. 3ª edición. Thomson. Belmont. USA. 2007.

ReFIP Matemática

RECURSOS PARA LA FORMACIÓN INICIAL DE PROFESORES DE EDUCACIÓN BÁSICA

La colección ReFIP es una serie de cuatro textos: Números, Geometría, Álgebra y Datos y azar, enfocados en la matemática para enseñar que requieren los profesores de Educación Básica.

Esta colección fue desarrollada en el proyecto FONDEF-D09I1023 “Recursos para la Formación Inicial de Profesores de Educación Básica en Matemática”, por un equipo de expertos disciplinarios y en educación de distintas universidades, liderados desde el Laboratorio de Educación del Centro de Modelamiento Matemático de la Universidad de Chile.

El proceso de elaboración de estos textos se llevó a cabo durante tres años y contempló el pilotaje de versiones preliminares en cursos de carreras de Pedagogía en Educación Básica de 16 universidades, en el que participaron alrededor de 5.000 estudiantes de Pedagogía de todo el país. Esto permitió hacer los cambios y ajustes necesarios para producir las versiones finales, y hacer que estos textos se constituyan en herramientas de gran utilidad en la formación docente.

Los textos promueven la reflexión acerca de la matemática escolar y su enseñanza, contribuyen a integrar conocimientos disciplinarios y pedagógicos, y tienen su foco en la matemática específica de la tarea de enseñar.

Más información acerca de la colección y el proyecto se encuentra en: **http://refip.cmm.uchile.cl/**

Álgebra
PARA FUTUROS PROFESORES DE EDUCACIÓN BÁSICA

Geometría
PARA FUTUROS PROFESORES DE EDUCACIÓN BÁSICA

Números
PARA FUTUROS PROFESORES DE EDUCACIÓN BÁSICA

Datos y azar
PARA FUTUROS PROFESORES DE EDUCACIÓN BÁSICA

9 789563 496178

CMM Center for Mathematical Modeling
Fondef FONDO DE FOMENTO AL DESARROLLO CIENTÍFICO Y TECNOLÓGICO
Fundación Luksic
Academia Chilena de Ciencias Instituto de Chile
sm
Ministerio de Educación Gobierno de Chile