





## Introducción a la Estadística

**ISBN:** 978-956-306-077-5

**Registro de Propiedad Intelectual:** 200.526

**Colección:** Herramientas para la formación de profesores de matemáticas.

**Diseño:** Jessica Jure de la Cerdá

**Diseño de Ilustraciones:** Catalina Frávega Thomas, Cristina Felmer Plominsky, Aurora Muñoz Lacourly

**Diagramación:** Pedro Montealegre Barba, Francisco Santibáñez Palma

**Financiamiento:** Proyecto Fondef D05I-10211

**Datos de contacto para la adquisición de los libros:**

**Para Chile:**

1. En librerías para clientes directos.
2. Instituciones privadas directamente con:

**Juan Carlos Sáez C.**

Director Gerente

Comunicaciones Noreste Ltda.

J.C. Sáez Editor

jcsaezc@vtr.net

[www.jcsaezeditor.blogspot.com](http://www.jcsaezeditor.blogspot.com)

Oficina: (56 2) 3260104 - (56 2) 3253148

3. Instituciones públicas o fiscales: [www.chilecompra.cl](http://www.chilecompra.cl)

**Desde el extranjero:**

1. Liberalia Ediciones: [www.liberalia.cl](http://www.liberalia.cl)
2. Librería Antártica: [www.antartica.cl](http://www.antartica.cl)
3. Argentina: Ediciones Manantial: [www.emanantial.com.ar](http://www.emanantial.com.ar)
4. Colombia: Editorial Siglo del Hombre  
Fono: (571) 3377700
5. España: Tarahumara, [tarahumara@tarahumaralibros.com](mailto:tarahumara@tarahumaralibros.com)  
Fono: (34 91) 3656221
6. México: Alejandría Distribución Bibliográfica, [alejandria@alejandrialibros.com.mx](mailto:alejandria@alejandrialibros.com.mx)  
Fono: (52 5) 556161319 - (52 5) 6167509
7. Perú: Librería La Familia, Avenida República de Chile # 661
8. Uruguay: Dolmen Ediciones del Uruguay  
Fono: 00-598-2-7124857

Introducción a la Estadística | Nancy Lacourly

Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile

[nlacourl@dim.uchile.cl](mailto:nlacourl@dim.uchile.cl)

**ESTA PRIMERA EDICIÓN DE 2.000 EJEMPLARES**

Se terminó de imprimir en febrero de 2011 en WORLDCOLOR CHILE S.A.

Derechos exclusivos reservados para todos los países. Prohibida su reproducción total o parcial, para uso privado o colectivo, en cualquier medio impreso o electrónico, de acuerdo a las leyes N°17.336 y 18.443 de 1985  
(Propiedad intelectual). Impreso en Chile.

# INTRODUCCIÓN A LA ESTADÍSTICA

Nancy Lacourly

Universidad de Chile





## **Editores**



**Patricio Felmer**, Universidad de Chile.  
Doctor en Matemáticas, Universidad de Wisconsin-Madison,  
Estados Unidos

**Salomé Martínez**, Universidad de Chile.  
Doctora en Matemáticas, Universidad de Minnesota,  
Estados Unidos

## **Comité Editorial Monografías**

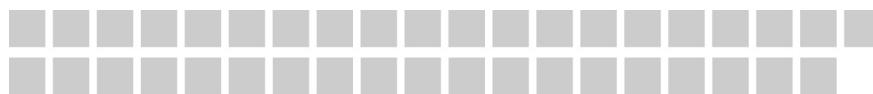


**Rafael Benguria**, Pontificia Universidad Católica de Chile.  
Doctor en Física, Universidad de Princeton,  
Estados Unidos

**Servet Martínez**, Universidad de Chile.  
Doctor en Matemáticas, Universidad de Paris VI,  
Francia

**Fidel Oteiza**, Universidad de Santiago de Chile.  
Doctor en Currículum e Instrucción, Universidad del Estado de Pennsylvania,  
Estados Unidos

**Dirección del Proyecto Fondef D05I-10211**  
**Herramientas para la Formación de Profesores de Matemática**



**Patricio Felmer**, Director del Proyecto  
Universidad de Chile.

**Leonor Varas**, Directora Adjunta del Proyecto  
Universidad de Chile.

**Salomé Martínez**, Subdirectora de Monografías  
Universidad de Chile.

**Cristián Reyes**, Subdirector de Estudio de Casos  
Universidad de Chile.

## Presentación de la Colección



La colección de monografías que presentamos es el resultado del generoso esfuerzo de los autores, quienes han dedicado su tiempo y conocimiento a la tarea de escribir un texto de matemática. Pero este esfuerzo y generosidad no se encuentra plenamente representado en esta labor, sino que también en la enorme capacidad de aprendizaje que debieron mostrar, para entender y comprender las motivaciones y necesidades de los lectores: Futuros profesores de matemática.

Los autores, encantados una y otra vez por la matemática, sus abstracciones y aplicaciones, enfrentaron la tarea de buscar la mejor manera de traspasar ese encanto a un futuro profesor de matemática. Éste también se encanta y vibra con la matemática, pero además se apasiona con la posibilidad de explicarla, enseñarla y entregarla a los jóvenes estudiantes secundarios. Si la tarea parecía fácil en un comienzo, esta segunda dimensión puso al autor, matemático de profesión, un tremendo desafío. Tuvo que salir de su oficina a escuchar a los estudiantes de pedagogía, a los profesores, a los formadores de profesores y a sus pares. Tuvo que recibir críticas, someterse a la opinión de otros y reescribir una y otra vez su texto. Capítulos enteros resultaban inadecuados, el orden de los contenidos y de los ejemplos era inapropiado, se hacía necesario escribir una nueva versión y otra más. Conversaron con otros autores, escucharon sus opiniones, sostuvieron reuniones con los editores. Escuchar a los estudiantes de pedagogía significó, en muchos casos, realizar eventos de acercamiento, desarrollar cursos en base a la monografía, o formar parte de cursos ya establecidos. Es así que estas monografías recogen la experiencia de los autores y del equipo del proyecto, y también de formadores de profesores y estudiantes de pedagogía. Ellas son el fruto de un esfuerzo consciente y deliberado de acercamiento, de apertura de caminos, de despliegue de puentes entre mundos, muchas veces, separados por falta de comunicación y cuya unión es vital para el progreso de nuestra educación.

La colección de monografías que presentamos comprende una porción importante de los temas que usualmente encontramos en los currículos de formación de profesores de matemática de enseñanza media, pero en ningún caso pretende ser exhaustiva. Del mismo modo, se incorporan temas que sugieren nuevas formas de abordar los contenidos, con énfasis en una matemática más pertinente para el futuro profesor, la que difiere en su enfoque de la matemática para un ingeniero o para un licenciado en matemática, por ejemplo. El formato de monografía, que aborda temas específicos

con extensión moderada, les da flexibilidad para que sean usadas de muy diversas maneras, ya sea como texto de un curso, material complementario, documento básico de un seminario, tema de memoria y también como lectura personal. Su utilidad ciertamente va más allá de las aulas universitarias, pues esta colección puede convertirse en la base de una biblioteca personal del futuro profesor o profesora, puede ser usada como material de consulta por profesores en ejercicio y como texto en cursos de especialización y post-títulos. Esta colección de monografías puede ser usada en concepciones curriculares muy distintas. Es, en suma, una herramienta nueva y valiosa, que a partir de ahora estará a disposición de estudiantes de pedagogía en matemática, formadores de profesores y profesores en ejercicio.

El momento en que esta colección de monografías fue concebida, hace cuatro años, no es casual. Nuestro interés por la creación de herramientas que contribuyan a la formación de profesores de matemática coincide con un acercamiento entre matemáticos y formadores de profesores que ha estado ocurriendo en Chile y en otros lugares del mundo. Nuestra motivación nace a partir de una creciente preocupación en todos los niveles de la sociedad, que ha ido abriendo paso a una demanda social y a un interés nacional por la calidad de la educación, expresada de muy diversas formas. Esta preocupación y nuestro interés encontró eco inmediato en un grupo de matemáticos, inicialmente de la Universidad de Chile, pero que muy rápidamente fue involucrando a matemáticos de la Pontificia Universidad Católica de Chile, de la Universidad de Concepción, de la Universidad Andrés Bello, de la Universidad Federico Santa María, de la Universidad Adolfo Ibáñez, de la Universidad de La Serena y también de la Universidad de la República de Uruguay y de la Universidad de Colorado de Estados Unidos.

La matemática ha adquirido un rol central en la sociedad actual, siendo un pilar fundamental que sustenta el desarrollo en sus diversas expresiones. Constituye el ciimiento creciente de todas las disciplinas científicas, de sus aplicaciones en la tecnología y es clave en las habilidades básicas para la vida. Es así que la matemática actualmente se encuentra en el corazón del currículo escolar en el mundo y en particular en Chile. No es posible que un país que pretenda lograr un desarrollo que involucre a toda la sociedad, descuide el cultivo de la matemática o la formación de quienes tienen la misión de traspasar de generación en generación los conocimientos que la sociedad ha acumulado a lo largo de su historia.

Nuestro país vive cambios importantes en educación. Se ha llegado a la convicción que la formación de profesores es la base que nos permitirá generar los cambios cualitativos en calidad que nuestra sociedad ha impuesto. Conscientes de que la tarea formativa de los profesores de matemática y de las futuras generaciones de jóvenes es extremadamente compleja, debido a que confluyen un sinnúmero de factores y disciplinas, a través de esta colección de monografías, sus editores, autores y todos los que han participado del proyecto en cada una de sus etapas, contribuyen a esta tarea, poniendo a disposición una herramienta adicional que ahora debe tomar vida propia en los formadores, estudiantes, futuros profesores y jóvenes de nuestro país.

**Patricio Felmer y Salomé Martínez**  
**Editores**



## Agradecimientos



Agradecemos a todos quienes han hecho posible la realización de este proyecto Fondef: "Herramientas para la formación de Profesores de Matemáticas". A Cristián Cox, quien apoyó con decisión la idea original y contribuyó de manera crucial para obtener la participación del Ministerio de Educación como institución asociada. Agradecemos a Carlos Eugenio Beca por su apoyo durante toda la realización del proyecto. A Rafael Correa, Edgar Kausel y Juan Carlos Sáez, miembros del Comité Directivo. Agradecemos a Rafael Benguria, Servet Martínez y Fidel Oteiza, miembros del Comité Editorial de la colección, quienes realizaron valiosos aportes a los textos. A Guillermo Marshall, Decano de la Facultad de Matemáticas de la Pontificia Universidad Católica de Chile y José Sánchez, entonces Decano de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Concepción, quienes contribuyeron de manera decisiva a lograr la integridad de la colección de 15 monografías. A Jaime San Martín, director del Centro de Modelamiento Matemático por su apoyo durante toda la realización del proyecto. Agradecemos a Víctor Campos, Ejecutivo de Proyectos de Fondef, por su colaboración y ayuda en las distintas etapas del proyecto.

Agradecemos también a Bárbara Ossandón de la Universidad de Santiago, a Jorge Ávila de la Universidad Católica Silva Henríquez, a Víctor Díaz de la Universidad de Magallanes, a Patricio Canelo de la Universidad de Playa Ancha en San Felipe y a Osvaldo Venegas y Silvia Vidal de la Universidad Católica de Temuco, quienes hicieron posible las visitas que realizamos a las carreras de pedagogía en matemática. Agradecemos a todos los evaluadores, alumnos, académicos y profesores -cuyos nombres no incluimos por ser más de una centena- quienes entregaron sugerencias, críticas y comentarios a los autores, que ayudaron a enriquecer cada uno de los textos.

Agradecemos a Marcela Lizana por su impecable aporte en todas las labores administrativas del proyecto, a Aldo Muzio por su colaboración en la etapa de evaluación, y también a Anyel Alfaro por sus contribuciones en la etapa final del proyecto y en la difusión de los logros alcanzados.

**Dirección del Proyecto**



## Índice General



<b>Prefacio</b>	19
<b>Capítulo 1: El pensamiento estadístico</b>	21
1.1 Un poco de historia	21
1.2 Probabilidad y estadística	22
1.3 Etapas de un estudio estadístico	23
1.4 Estadística descriptiva y estadística inferencial	25
1.5 Ejemplos	25
1.6 Ejercicios	27
<b>Capítulo 2: Estadística descriptiva: cómo hacer hablar los datos</b>	29
2.1 Introducción	29
2.2 Población, individuos y variables	31
2.3 Tipos de variables	32
2.4 Distribución de frecuencias y sus representaciones gráficas	33
2.5 Resúmenes de la distribución de frecuencias	43
2.6 Diagrama de caja y el resumen de cinco números	52
2.7 Cómo visualizar dos o más variables al mismo tiempo	57
2.8 Resumen de la terminología	61
2.9 Ejercicios	61
<b>Capítulo 3. Introducción a la inferencia estadística</b>	71
3.1 Motivación	71
3.2 Teoría de muestreo	73
3.3 Distribución en la población y distribución en la muestra	78
3.4 Distribuciones teóricas	84
3.5 Intervalos de confianza	90
3.6 Tests de hipótesis	95
3.7 Resumen de la terminología	107
3.8 Ejercicios	108
<b>Capítulo 4. Regresión lineal simple</b>	113
4.1 Introducción a los modelos	113

4.2 Coeficiente de correlación lineal	115
4.3 Planteamiento de la regresión lineal	119
4.4 Criterio de mínimos cuadrados	120
4.5 Predicciones	124
4.6 Validación del modelo	124
4.7 Regresión de $Y$ sobre $X$ y regresión de $X$ sobre $Y$	126
4.8 Resumen de la terminología	127
4.9 Ejercicios	127
<b>Solución de los ejercicios</b>	131
<b>Tablas estadísticas</b>	145
<b>Bibliografía</b>	149
<b>Índice de Figuras</b>	151
<b>Índice de nombres propios</b>	153
<b>Índice de Términos</b>	155

*La estadística es la gramática de la ciencia*  
Karl Pearson



## Prefacio



La introducción de las Probabilidades y Estadística en la Enseñanza Media no ha sido fácil para los profesores de Matemática. Es posible que la dificultad emane, precisamente, de una formación basada en la concepción de la matemática como una “ciencia exacta”, reducible en último término a la aplicación de algoritmos. En esta monografía intentamos introducir al lector al pensamiento estadístico, sin reducirlo al enunciado ciego de recetas, puesto que éste es una necesidad para la cultura científica de los ciudadanos. *El pensamiento estadístico nace cuando se toma conciencia de la existencia de la fluctuación de las muestras*<sup>1</sup>.

Nuestro mundo actual está lleno de imágenes. ¿Quién no ha visto en la prensa gráficos que muestran los resultados de una elección, los cambios económicos del país, etc.? Los medios informativos nos bombardean diariamente con los resultados de encuestas de opinión, en donde se nos habla de muestreo y se nos presentan gráficos y cifras. Si no tenemos nociones de base en estadística, seremos incapaces de evaluar el contenido de estas informaciones.

Los gráficos, que muestran resultados estadísticos, pueden ser complejos. Es importante elegir el gráfico más adecuado para una situación dada, construirlo correctamente y saber interpretarlo. Veremos por qué se usan las muestras y cuánto podemos creer en los resultados de una encuesta que sólo pudo entrevistar a una parte de la población.

Los conceptos estadísticos no son fáciles de enseñar en el colegio, especialmente la noción de azar, pero constituyen un excelente terreno para actividades interdisciplinarias. Veremos distintas maneras de “hacer hablar” los datos obtenidos de encuestas o experimentos. Discutiremos las herramientas de estadística descriptiva que, a pesar de ser más familiares para la mayoría de la gente, son a menudo delicadas de interpretar. Reflexionaremos sobre conceptos básicos como el azar, la variabilidad de los resultados obtenidos de una muestra, el modelamiento probabilístico de los problemas de inferencia, la noción de riesgo en la toma de decisiones y el uso de modelos matemáticos para hacer predicciones.

En esta monografía, varios resultados no se demuestran, prefiriendo concentrarse en explicar los conceptos de la estadística y en la interpretación de los resultados.

---

<sup>1</sup>Commission Permanente de Réflexion sur l’Enseignement des Mathématiques (Direction des Lycées), 21 Octubre 1983, Francia.

El contenido de los distintos capítulos es el siguiente: En el Capítulo 1 comentamos los conceptos generales y el modo de pensar de la Estadística; en el Capítulo 2, presentamos la Estadística Descriptiva, a veces llamada también *Análisis exploratorio de datos*; en el Capítulo 3 entregamos los elementos básicos de la Inferencia Estadística; en el Capítulo 4 introducimos los modelos estadísticos predictivos a través de la regresión lineal simple. Hemos intercalado referencias históricas cuando eso nos pareció relevante, y agregado ejercicios de autoevaluación para ayudar a la comprensión del texto, además de ejercicios que pueden utilizarse con los estudiantes de Enseñanza Media (estos ejercicios se marcarán con una \*). La solución de los ejercicios se encuentran en un anexo junto con algunas referencias bibliográficas y sugerencias de lectura adicional.

Antes de terminar, expreso mi sincero agradecimiento a Lorena Cerdá que tuvo la gentileza de dedicar parte de su valioso tiempo en la revisión de esta monografía.

Quiero agradecer especialmente a los editores, Patricio Felmer y Salomé Martínez, no sólo por haber desarrollado el proyecto Fondef, que hizo posible esta monografía, sino además por la enorme dedicación que aplicaron a su corrección y producción. Agradezco también a Francisco Santibáñez por todo el tiempo dedicado a la diagramación del texto para su mejoramiento.

Finalmente dedico este trabajo a Juan Muñoz, mi esposo, quien siempre me prestó apoyo y sabe lo importante que ha sido para mí escribirlo. Además de revisarlo, puso su grano de arena con las largas discusiones que tuvimos.

## Capítulo 1: El pensamiento estadístico



Desde los primeros cursos de matemáticas, los niños aprenden a trabajar con formalismos, estructuras totalmente definidas, y con la noción de que las propiedades sólo pueden ser ciertas o falsas. No hay que dejar nada al azar. El pensamiento estadístico no se opone a esas ideas, pero las maneja en forma más atrevida: el azar y el riesgo controlado pasan a ser herramientas fundamentales.

La Estadística es una rama del método científico que se ocupa del manejo de datos empíricos, es decir, de datos obtenidos al contar o medir fenómenos naturales, con instrumentos y métodos que generalmente entregan resultados *inciertos*. Además ofrece métodos para la recolección, la agregación y el análisis de esos datos.

### 1.1 Un poco de historia

Desde el principio de la historia, los estados se preocuparon por contar y medir sus dominios y sus súbditos. Así lo hicieron, por ejemplo, los chinos, los egipcios y los incas. Por muchos años, se miraban esos *datos estadísticos* como un reflejo preciso de la realidad, pero a partir del siglo XIX, la *estadística inferencial* comenzó a penetrar en el dominio de lo incierto.

La estadística inferencial surgió del concepto de azar, que es tan antiguo como los juegos (los dados y los juegos con huesos que, en Chile llamamos "payayas", son antíquissimos) y motivó desde antaño las reflexiones de los filósofos.

Durante la Edad Media hubo una gran actividad científica y artística en Oriente y el nombre de azar parece haberse trasladado desde Siria a Europa. La flor de azahar, que aparecía en los dados de la época, podría ser el origen de la palabra. Desde tiempos muy antiguos existen las compañías aseguradoras que iniciaron investigaciones matemáticas sobre este tema.

En el siglo XVII aparecieron los primeros problemas famosos de juegos de azar. Las reglas de cálculo, desarrolladas hasta entonces para estos juegos, se vieron aplicadas a otras disciplinas. Los censos demográficos, que se hacían desde la antigüedad, requieren recolectar muchos datos. En Inglaterra, a pesar que John Graunt (1620-1674) ya la había introducido algunos años antes, fue Edmund Halley (1656-1742), más conocido por el cometa que lleva su nombre, quien construyó por primera vez una *tabla de mortalidad*, que mostraba la cantidad de decesos observados por causa. El nacimiento de la Estadística podría situarse en el paso de una *lista de observaciones* a una *tabla*, que muestre sumas, frecuencias, promedios, etc. Con Cristiaan

Huygens (1629-1695) aparecen los primeros gráficos. En 1669, presentó las tablas de mortalidad en forma gráfica, mostrando el número de sobrevivientes en función de la edad. La demografía y los seguros de vida aprovecharon el desarrollo de la teoría de las probabilidades. Por ejemplo, los sexos de una sucesión de niños recién nacidos se pueden representar como lanzamientos repetidos de una moneda, con niños y niñas, en vez de caras y sellos. De la misma manera, podemos considerar la proporción de hombres de 50 años que van a vivir un año más, haciendo una analogía con seguir vivo (“cara”) o morir (“sello”).

Durante el siglo XVIII, Pierre Simón Marqués de Laplace (1749-1827) pasó de la observación estadística a la creación de conceptos probabilísticos, asimilando estos problemas a juegos, y basándose en la frecuencia relativa de los eventos para definir la probabilidad que nazca una niña, o que un hombre de 50 muera en el año.

Si bien la extensión de los juegos de azar a la demografía o a la matemática actuaria fue extremadamente importante, los planteamientos iniciales se vieron limitados por considerar todos los resultados posibles como simétricos (equiprobables). Por ejemplo, Daniel Bernoulli (1700-1782), careciendo de datos sobre la mortalidad por viruela en distintas edades, supuso que el riesgo de morir de la enfermedad era el mismo en todas ellas. Lo que evidentemente es muy discutible.

Actualmente todos los gobiernos recogen sistemáticamente datos sobre su población, su economía, sus recursos naturales y su condición política y social para tomar decisiones. En las actividades industriales o comerciales, las estadísticas son parte de la organización. Lo mismo sucede en los sectores agrícolas y forestales, cuando se requiere predecir la producción. En la investigación científica (medicina, física, biología, ciencias sociales, etc.) el rol de la estadística es primordial.

En materias de educación, por ejemplo, un psicólogo puede medir las aptitudes intelectuales de un grupo de estudiantes y darles un método de estudio. El rendimiento de los estudiantes permitirá evaluar el método en función de las aptitudes. La psicométría es la rama de la psicología que se ocupa de medir, mediante tests llevados a escalas numéricas, características psicológicas relativas al comportamiento, el aprendizaje y el rendimiento de los individuos.



## 1.2 Probabilidad y Estadística

En esta sección mostraremos cómo se razona en Estadística y en qué medida el modo de pensar de ella difiere del de la matemática.

Para mostrar cómo se razona en Estadística, haremos una comparación con el razonamiento utilizado en Probabilidad, que es matemático, apoyándonos en un ejemplo.

En teoría de probabilidad, para deducir que la probabilidad de que al lanzar un dado salga un número impar es igual a  $1/2$ , se parte suponiendo que si, el dado no está cargado, los números 1 a 6 son equiprobables; o sea, a partir de un modelo probabilístico adecuado (en este caso, el supuesto de equiprobabilidad), se deducen nuevos modelos o propiedades. En Estadística, en cambio, tratamos de responder, por ejemplo, a la pregunta *¿el dado está cargado?*, comprobando si el modelo probabilístico de equiprobabilidad subyacente está en acuerdo con los datos experimentales obtenidos al lanzar el dado muchas veces.

En Estadística, se tiene un fenómeno a estudiar o una *hipótesis de trabajo*: “*El dado no está cargado*”. Para verificar la hipótesis, se realiza un **experimento**; por ejemplo, lanzar 100 veces el dado. De los porcentajes registrados para cada uno de los resultados (los números del 1 al 6) se tratará de concluir si el dado está cargado.

No hay que confundir el uso de la palabra “**estadísticas**” (plural), que designa un conjunto de datos observados y la palabra “**estadística**” (singular), que es la rama del método científico que se ocupa de estos datos observados. En el título de esta monografía la palabra está en singular.

### 1.3 Etapas de un estudio estadístico

El conjunto total de los objetos que se quiere estudiar (personas, hogares, niños de 2 a 5 años, flores, etc.) se llama *población*. Una muestra es una parte de los elementos de la población, y de ella se extrae información que permite entonces decir algo sobre la población entera. Para que la información de la muestra permita decir cosas útiles sobre la población, no se puede sacar la muestra de cualquier manera y es necesario saber cómo se obtuvo. En otras palabras, cuál fue el *diseño muestral* y cómo se puso en práctica.

Un estudio estadístico se desarrolla en varias etapas:

- definir el problema a estudiar, especialmente los objetos de interés;
- elegir un experimento o una muestra al azar entre los objetos del estudio, cuando no se trata de un censo;
- recolectar los datos;
- a partir de un estudio matemático de los datos recolectados, resumirlos con el riesgo de perder información, pero con ello, ganar en interpretación. Esta es la llamada **Estadística Descriptiva**, que trata de resaltar lo fundamental de los datos eliminando *el ruido*. En el medio ambiente, se define como ruido todo sonido no deseado por el receptor. En estadística, se llama *ruido* a variaciones que perturban el fenómeno de interés. Son variaciones cuya información no es identificable claramente. Los errores de mediciones son, en general, ruido;

- deducir resultados para la población y tomar decisiones en base a estos, aceptando ciertos riesgos supuestamente controlados. Esta es la llamada **Estadística Inferencial**.

### 1.3.1 Recolección de los datos

Hay que distinguir los censos de los muestreos. Los censos recogen datos sobre la totalidad de la población considerada; los muestreos sólo lo hacen sobre una parte de la población. *¿Cómo elegir una muestra de una población para obtener informaciones fidedignas sobre esta población?* La forma de elegir la muestra puede ser muy compleja, pero generalmente eso se hace en forma aleatoria, lo que conduce a aplicar la teoría de las probabilidades [12].

### 1.3.2 Descripción estadística de los datos

La descripción estadística permite resumir, reducir y presentar gráficamente el contenido de los datos, con el objeto de facilitar su interpretación sin preocuparse de si éstos provienen de una muestra o no. Las técnicas utilizadas dependerán de la cantidad de datos, de la cantidad de variables, de la naturaleza de los datos y de los objetivos del estudio. Esta etapa del estudio es una ayuda para el análisis inferencial.

### 1.3.3 Análisis inferencial

El análisis inferencial es la etapa más importante del razonamiento estadístico, el cual está basado en un modelo matemático o probabilístico.

**La inferencia estadística** consiste en establecer conclusiones sobre la población a partir de las características de una muestra. Se basa en modelos que dependen de los objetivos del estudio, de los datos y del conocimiento *a priori* que se pueda tener sobre el fenómeno estudiado. El modelo no está en general totalmente determinado (es decir, se plantea una familia de modelos de un cierto tipo); por ejemplo, la familia de las distribuciones normales, la familia de las distribuciones de Poisson o un modelo lineal. Estos modelos tendrán algunos elementos indeterminados llamados **parámetros**. Se trata entonces de precisar lo mejor posible tales parámetros desconocidos, a partir de los datos empíricos observados en la muestra: **esto se llama estimación estadística**. Por otro lado, antes o durante el análisis, se tienen generalmente consideraciones teóricas respecto del problema estudiado y se trata entonces de comprobarlas o rechazarlas a partir de los datos obtenidos en la muestra: **Esto se consigue aplicando los llamados tests estadísticos**.

Por ejemplo, se quiere estudiar la calidad de una partida de fósforos entregada por un cierto fabricante. Obviamente, no podemos encenderlos todos para sacar una conclusión. Observamos entonces una muestra de, por ejemplo, 500 fósforos. Nos preguntamos entonces:

- *¿Cómo seleccionar los 500 fósforos?*
- *¿Cómo extraer o inferir las conclusiones obtenidas sobre la muestra de 500 fósforos a la totalidad de los fósforos de la partida?*

La teoría del muestreo responde a la primera pregunta y la inferencia estadística a la segunda.

### 1.3.4 Decisión o predicción

El análisis está condicionado por la finalidad del estudio, y generalmente consiste en tomar una decisión o hacer un pronóstico. Por ejemplo, decidir si una marca de ampolletas está conforme a las normas de calidad (duración mínima de 2500 horas) o si un tratamiento es eficaz para combatir la hipertensión, son **problemas de decisión** bastante comunes. Predecir el IPC del próximo mes, las temperaturas mínima y máxima de mañana en Santiago o el porcentaje de votos de un candidato en una elección, a partir de algunas muestras, son **problemas de predicción**.

## 1.4 Estadística Descriptiva y Estadística Inferencial

Antes que nada, es necesario distinguir la Estadística Descriptiva de la Estadística Inferencial. Las herramientas de la primera, tratan de mostrar los datos, provenientes de una muestra o un censo, para ayudar a su interpretación, sin hacer ninguna hipótesis del fenómeno que pretenden representar. Con las herramientas de la segunda, tratamos, en cambio, de obtener conclusiones más allá de los datos observados, haciendo, en general, hipótesis sobre los datos. Hacemos inferencia, por ejemplo, cuando intentamos predecir el resultado de una elección a partir de una encuesta de opinión obtenida sobre una muestra de votantes. También hacemos inferencia para decidir si existe una diferencia de sueldos observada entre hombres y mujeres para un mismo trabajo: Es un fenómeno de validez global y no un resultado casual de la muestra de datos utilizada para el estudio. La Inferencia Estadística permite sacar conclusiones a partir de datos observados en una muestra para toda la población, mientras que la Estadística Descriptiva muestra simplemente lo que hay en los datos.

Más formalmente, podemos decir que la **Estadística Descriptiva** consiste en resumir y representar informaciones, mientras que la **Inferencia Estadística**, consiste en sacar resultados sobre una muestra, para inferir conclusiones sobre la población de donde ésta proviene.

Los estudios estadísticos generalmente utilizan ambos métodos.

## 1.5 Ejemplos

**Ejemplo 1.1.** Un sociólogo quiere determinar el ingreso anual promedio de los hogares de la Región Metropolitana. Como recolectar esta información para todos los hogares sería prohibitivo, el sociólogo decide usar una muestra. Puede hacer esto porque no le interesa conocer el ingreso anual de cada familia en particular, sino el ingreso anual promedio de la totalidad de los hogares que viven en la Región Metropolitana, y eventualmente, la distribución de estos ingresos en la población.

**Ejemplo 1.2.** Para saber cuál es el número total  $N$  de pejerreyes del lago Rapel, sería inconcebible pescarlos todos. Se pueden, entonces, pescar algunos,  $A = 200$

por ejemplo, marcarlos y devolverlos al lago. Volver a pescar al otro día,  $n = 100$  por ejemplo, y observar el número  $k$  de marcados en la segunda muestra. Por este procedimiento se puede, entonces, estimar el número total  $N$  de peces en el lago, suponiendo que la proporción de peces marcados en el lago y la proporción de peces marcados en la muestra son iguales:

$$\frac{A}{N} = \frac{k}{n} \Rightarrow N = \frac{n}{k} A$$

Si se encontró  $k = 16$  peces marcados en la segunda muestra de  $n = 100$  peces, se estimará que hay  $N = \frac{100}{16} \times 200 = 1250$  en el lago.

**Ejemplo 1.3.** Un candidato a una elección presidencial encarga a un centro de estudios de opinión un análisis sobre el porcentaje de votos que podría obtener en la elección que tendrá lugar en un mes más. El centro de estudios realiza un sondeo de opiniones sobre 1500 personas elegidas al azar en la población de votantes. Luego, le informa al candidato que si la elección tuviera lugar ese mismo día tendría 45 % de votos contra 55 % de su adversario. Además, señala que esta estimación tiene un error de 2,52 % con un nivel de confianza de 95%<sup>1</sup>. Con este pronóstico, el candidato concluye que debe cambiar su estrategia de campaña electoral.

**Ejemplo 1.4.** El Ministerio de Educación quiere analizar la brecha entre colegios municipales y particulares pagados a través de los resultados de la prueba SIMCE del 2do medio. En este caso, la población está constituida por los colegios de Enseñanza Media. El problema no es inferir los resultados desde una muestra del país (pues la información de los resultados del SIMCE están disponibles para todos los colegios), sino estudiar si las diferencias entre colegios privados pagados y colegios municipales son *significativas*.

**Ejemplo 1.5.** La Universidad de Chile quiere saber si la PSU es un buen predictor del rendimiento futuro de los estudiantes que ingresan a la Universidad.

Los problemas citados son muy distintos; algunos se basan en datos censales y otros en datos muestrales. Pero hay ciertos elementos y una línea general de razonamiento que le son comunes. Se trata de responder preguntas a partir de los datos empíricos que describen un fenómeno.

Cuando no se puede usar un censo, el problema es cómo elegir una muestra para poder sacar conclusiones que sean válidas para la población entera. Cada individuo o elemento de la muestra sólo interesa en la medida en que es parte de la población. La teoría de muestreo ofrece métodos para obtener muestras.

---

<sup>1</sup>Se explicará el sentido estadístico exacto de esta frase en la página 90

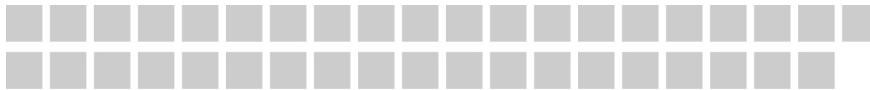
## 1.6 Ejercicios

Los ejercicios siguientes pueden utilizarse en los estudios de Enseñanza Media.

1. En cada ejemplo citado anteriormente, especifique:
  - (a) La población;
  - (b) Si el estudio es censal o muestral.
  - (c) ¿Cuál es la pregunta que se quiere responder?
2. Revise su compresión de las ideas fundamentales:
  - (a) ¿Cuál es el sentido de las palabras Estadística y Estadísticas?
  - (b) ¿Cuál es la importancia de las Estadísticas para un Estado?
  - (c) ¿Para qué le puede servir la Estadística a una industria?
3.
  - (a) ¿Con qué periodicidad se hace el censo poblacional en Chile?
  - (b) ¿Cómo se hace el censo?
  - (c) ¿En qué difiere el censo poblacional y la encuesta CASEN en Chile?
4.
  - (a) ¿Qué diferencia existe entre la Estadística Descriptiva y la Estadística Inferencial?
  - (b) ¿Por qué se usa el Cálculo de Probabilidades en Estadística?



## Capítulo 2: Estadística descriptiva: cómo hacer hablar los datos



### 2.1 Introducción

La estadística descriptiva tiene su origen mil o dos mil años antes de Cristo, en Egipto, China y Mesopotamia, donde se hacían censos<sup>1</sup> para la administración de los imperios. Los egipcios tuvieron el barómetro económico más antiguo: un instrumento llamado *nilómetro*, que medía el caudal del Nilo y servía para definir un índice de fertilidad, a partir del cual se fijaba el monto de los impuestos. Con la variabilidad del clima, ya conocían el concepto de incertidumbre.

El quipu (en quechua significa *nudo*), fue un sistema nemotécnico de cuerdas de lana o algodón y nudos de colores, que los funcionarios del Imperio Inca utilizaban como sistema de contabilidad y registro de censos de población y de cosechas. Todavía no se conocen bien los códigos usados, ni se explotan los datos recopilados.

Como muestra, esta es parte de la Tabla de Mortalidad de la ciudad de Londres del año 1632	
Bautizados: Varones (4594); Hembras (4590)	
Enterrados: Varones (4932); Hembras (4603) De los cuales 8 por peste.	
• Abortos (445)	• Gusos (27)
• Afligidos (74)	• Hermonías y almorranas (1)
• Accidentes (16)	• Hidropesía y abatagamias (267)
• Aflicción (11)	• Higado desarrollado (187)
• Afta y dolor de boca (40)	• Ictericia (43)
• Ahogados (34)	• Indigestión (86)
• Ahogados o hambrientos en lactancia (7)	• Infantes (2268)
• Ancianos (628)	• Letargia (50)
• Anginas (7)	• Lunáticos (5)
• Apoplejía (17)	• Mordido por un perro rabioso (1)
• Asesinatos (7)	• Muertos en la calle y de hambre (6)
• Asma (1)	• Nauseas (7)
• Cáncer y lupus (10)	• Pestis (25)
• Chancro (1)	• Pestaña (8)
• Cláctica (1)	• Planeta (13)
• Cólico, cálculo y estrangurria (56)	• Pleuresia y bazo (36)
• Consultación (1797)	• Pustulas y viruelas (531)
• Convulsiones, pérdidas de sangre, llagas y úlceras (28)	• Quemaduras y escaldaduras (5)
• Convulsión (241)	• Relajación y temblores (9)
• Corte de cálculo (5)	• Repentiamamente (62)
• Depresión (8)	• Resfriado y tos (55)
• Desengarramiento (3)	• Sarampión (80)
• Dientes (470)	• Sobrepeso (171)
• Disentería y flujo de sangre (348)	• Sudorosa (15)
• Ejecutados y condenados (18)	• Tabardillo y fiebre maligna (38)
• Escorbuto y sarna (9)	• Timpanización (13)
• Escrófula (38)	• Tisis (34)
• Fiebre (108)	• Tumor de pulmones (98)
• Fiebre intermitente (43)	• Varicela (6)
• Fistula (13)	• Viruela (12)
• Gangrena (5)	• Vomitos (1)
• Gota (4)	

La **Estadística Descriptiva** es un conjunto de técnicas para construir tablas, resumir y representar gráficamente conjuntos de datos provenientes de un censo o una muestra, con el fin de mostrar sus aspectos más relevantes, permitir una mejor interpretación de los fenómenos estudiados y preparar los estudios inferenciales.

Sin embargo, sólo con las tablas de mortalidad de John Graunt (1620-1674) y los primeros gráficos de Cristiaan Huygens, que mostraban los sobrevivientes en función de la edad, vemos, en el siglo XVII, el inicio de la estadística descriptiva, que permite resumir los datos recopilados y ayudar a la predicción y la toma de decisión. La tabla adjunta muestra la mortalidad por causa en la ciudad de Londres, en el año 1632.

<sup>1</sup>La palabra censo viene de la palabra latina *censere* que significa fijar impuestos.

Es importante familiarizarse con estas herramientas para facilitar la comprensión de los fenómenos que se estudian y tomar así decisiones acertadas. Se pueden clasificar las herramientas descriptivas más usuales en tres tipos complementarios:

- Gráficos, que permiten visualizar la distribución de los datos.
- Tablas, que muestran los datos de manera más interpretable.
- Estadígrafos o estadísticos, que son números que resumen lo esencial de los datos.

En general, los datos estadísticos se presentan como listas de personas u objetos, con mediciones sobre cada uno de ellos.

**Ejemplo 2.1.** El profesor Cabello quiere estudiar el progreso en matemáticas de sus 15 alumnos de primero medio. En la Tabla 2.1, encontramos en la primera columna un identificador del alumno; las otras columnas indican el nombre, género, las notas de seis pruebas de matemáticas, el número de hermanos y una evaluación de la conducta (mala, regular, buena) de cada uno.

Nos preguntamos, por ejemplo:

- ¿Se parecen en algo las 6 notas de cada alumno?
- ¿Se parecen en algo las notas de los 15 alumnos en cada prueba?
- ¿Las 6 notas de los 15 alumnos varían de manera parecida?
- ¿Influye la conducta en el rendimiento en matemáticas?
- ¿Hay diferencia de rendimiento entre niños y niñas?
- ¿La mayoría de los alumnos tiene a lo más 2 hermanos?

TABLA 2.1. Curso del profesor Cabello

ID	Nombre	Género	Notas matemática						Nº hermanos	Conducta
			1	2	3	4	5	6		
1	Juan	Masculino	5,2	5,0	4,8	5,5	5,8	6,2	0	Mala
2	Pedro	Masculino	4,7	3,8	4,4	5,7	4,6	6,6	2	Regular
3	María	Femenino	3,2	4,0	4,9	6,2	5,1	6,5	2	Buena
4	Carmen	Femenino	4,8	4,6	4,8	5,1	5,0	5,2	1	Regular
5	Emilio	Masculino	5,8	5,6	5,9	5,8	6,4	6,2	3	Buena
6	Aurora	Femenino	6,8	6,2	6,9	6,8	6,4	6,5	3	Buena
7	Rodrigo	Masculino	5,8	4,6	5,2	6,6	5,4	5,5	1	Regular
8	Silvia	Femenino	3,8	3,6	3,2	4,6	3,4	4,5	0	Regular
9	Patricia	Femenino	3,4	3,8	4,2	4,6	5,4	4,5	2	Buena
10	Andrés	Masculino	2,4	3,8	3,2	3,6	2,4	4,5	4	Mala
11	Fran	Masculino	5,4	4,8	5,2	5,6	4,4	5,5	2	Buena
12	Roberto	Masculino	5,8	5,8	6,2	6,6	6,4	6,5	5	Regular
13	Inés	Femenino	6,0	6,4	6,2	6,5	6,6	6,4	1	Regular
14	Alberto	Masculino	3,9	4,5	3,2	5,0	4,6	5,4	3	Regular
15	Marcela	Femenino	4,9	5,4	4,2	5,5	5,6	5,9	2	Buena

**Ejemplo 2.2.** La Universidad de Chile quiere saber si la PSU es un buen predictor del rendimiento futuro del estudiante que ingresa en la Universidad. La Tabla 2.2 muestra los resultados de algunos de los 169,994 alumnos que rindieron la PSU en 2005. Por su tamaño, es imposible presentar toda la tabla aquí, lo que es una clara diferencia con la Tabla 2.1. Pero si no nos interesamos en los resultados de nuestro hijo u otro alumno en particular, sino sólo en comparar, por ejemplo, los rendimientos entre colegios municipales y privados, o entre géneros, podemos aplicar los métodos de la estadística descriptiva a los puntajes PSU, de manera de presentar los datos con el fin de ver si hay diferencias entre colegios.

TABLA 2.2. Algunos resultados de la PSU 2005

ID	NEM	LENG	MAT	H. CSOC	CIENC	DEP.	Colegio	RAMA
859408	5,0	252	377			Mu	L. LA PORTADA A22	H. CIENT. D.
854951	5,5	314	342	360		Mu	L. A. SABELLA	H. CIENT.D.
899343	5,4	291	414	297		PS	L. LA PORTADA A22	TEC. Y SERV.
853762	5,2	523	606		525	PS	L. M. BAHAMONDE SILVA	H. CIENT. N.
899528	6,3	724	524	583	569	Mu	L. A. SABELLA	H. CIENT. D.

**Ejemplo 2.3.** El Ministerio de Educación quiere analizar la brecha entre colegios municipales y particulares pagados a través de los resultados de la prueba SIMCE. Como en el ejemplo anterior, mostramos solamente algunos de los resultados promedio por colegio del SIMCE 2006 en 2do Medio (Tabla 2.3)<sup>2</sup>.

TABLA 2.3. Algunos resultados SIMCE 2do medio 2006

Colegio	Región	Comuna	Rural	Dep.	GSE	Nº leng	Prom leng	Nº Mat	Prom Mat
L. Poli Arica	1	Arica	U	MD	B	284	281	208	194
L. J. Rusque Portal	5	Nogales	U	MD	B	41	41	241	243
C. Adv. Maranata	4	La Serena	R	PS	C	114	115	252	232

Antes de describir las herramientas descriptivas de la Estadística, es importante definir algunos conceptos básicos: *población*, *individuo* y *variable*.

## 2.2 Población, individuos y variables

En las noticias escuchamos hablar cotidianamente de **población** para referirse a un grupo de personas que tienen algo en común, como la población de los chilenos o la población de los niños de Santiago. Para el estadístico, el concepto de **población** se refiere a un conjunto de elementos (personas, animales, plantas, objetos, colegios, etc.), sobre el que se quiere averiguar algo, a partir de mediciones, encuestas, etc., realizadas a sus integrantes. Este término proviene de las primeras aplicaciones de la Estadística, que se realizaron en el campo de la demografía. Los elementos que definen una población son de la misma naturaleza, en el sentido que tienen propiedades en

<sup>2</sup><http://www.simce.cl/>, <http://www.demre.cl>

común. Para hablar de los elementos que conforman la población usaremos en general las palabras **individuos** u **observaciones**.

La Estadística trata en general de las propiedades de la población más que de las propiedades de sus individuos en particular. En el ejemplo 2.1, nos interesamos en el número de alumnos con promedio mayor o igual a 4,0 o en la proporción de niñas en el curso.

En las estadísticas demográficas de Chile nos interesamos en la proporción de hombres mayores de 60 años y no de saber si Juan y Pedro son mayores de 60 años. La talla, la edad, el género o el color del pelo de una persona son mediciones o caracteres de la persona. En la población chilena, las personas no tienen todas la misma edad. Las características varían en general entre los individuos de la población. Para hablar de las características medidas u observadas en la población hablaremos de **variables o caracteres**.

Supondremos en este capítulo, que los datos fueron medidos sobre toda la población. El caso en que los datos sólo se conocen para una muestra, se puede aplicar las herramientas de este capítulo, sin preocuparse que se trata de una muestra.

En un estudio estadístico es importante determinar claramente la población y las variables a medir, los que dependen del objetivo de éste. En el ejemplo 2.1 la población está definida por los 15 alumnos del curso del profesor Cabello, donde cada fila corresponde a un individuo de la población - o sea, a un alumno - y cada columna corresponde a una variable. El número de identificación, el nombre, el género, los resultados de cada prueba, el número de hermanos y la conducta son las variables.

En el ejemplo 2.2, la población está formada por todos los alumnos que rindieron la PSU en el 2005 y los puntajes, la NEM, etc., son las variables. En el ejemplo 2.3, la población está formada por los colegios evaluados en el SIMCE 2do Medio en 2006. Aquí los colegios son los *individuos* y la región, la comuna, los puntajes, etc. son las variables.

Los estadísticos suelen organizar la información disponible en arreglos rectangulares como los que mostramos en las Tablas 2.1, 2.2 y 2.3, con los individuos en filas y las variables en columnas. Estos arreglos se llaman **tablas de datos** y pueden residir en hojas de papel o en medios informáticos como, por ejemplo, hojas Excel.

### 2.3 Tipos de variables

En los tres ejemplos previos, las variables son de distinta naturaleza. Algunas se definen con números decimales, otras con números enteros, y otras con características no numéricas, definidas por medio de textos, tales como el nombre, el género y la comuna.

Las variables pueden clasificarse según los valores o alternativas que pueden tomar. Consideremos el puntaje de matemática en la PSU. La población es el conjunto  $F$  de todos los alumnos que rindieron la PSU en el 2005 y el puntaje de matemáticas es una función  $X : F \rightarrow \mathbb{R}$ , donde  $\mathbb{R}$  es el conjunto de los números reales.

Si consideramos el nombre del colegio de los alumnos que rindieron la PSU, la población es la misma pero el conjunto de alternativas que puede tomar la variable *colegio* no son números, sino *nombres*.

Más generalmente una variable o carácter es una función

$$X : F \rightarrow \mathcal{Q}$$

donde  $F$  es la población y  $\mathcal{Q}$  es el conjunto de valores que puede tomar la variable.

Podemos distinguir cuatro tipos de variable según la naturaleza del conjunto  $\mathcal{Q}$ .

- Variables **continuas** (también llamadas **cuantitativas**), cuyos valores son números reales ( $\mathcal{Q} = \mathbb{R}$ ). Pueden ser mediciones físicas, fisiológicas, económicas, etc. Por ejemplo, el ingreso, la edad, el peso o la talla, una nota o puntaje en una prueba.
- Variables **discretas**, que toman valores en un conjunto finito de números, en general enteros. Por ejemplo, el número de hijos de una familia  $\mathcal{Q} = \{0, 1, 2, 3, \dots\}$ ; el número de dormitorios de un departamento  $\mathcal{Q} = \{1, 2, 3, \dots\}$ , el número de genes de un cromosoma.
- Variables **nominales** (o **cualitativas**), cuyos valores representan atributos no numéricos. Estos valores se llaman **categorías o modalidades**. Por ejemplo,
  - (i) el género,  $\mathcal{Q} = \{\text{hombre}, \text{mujer}\}$ ,
  - (ii) el estado civil,  $\mathcal{Q} = \{\text{soltero}, \text{casado}, \text{separado}, \text{viudo}\}$ ,
  - (iii) el color de los ojos de una persona,  $\mathcal{Q} = \{\text{azul}, \text{verde}, \text{café}, \text{negro}\}$ ,
  - (iv) el voto por un candidato,  $\mathcal{Q} = \{\text{Valverde}, \text{Rojas}\}$  .
- Variables **ordinales**, que son variables nominales cuyas categorías pueden ordenarse (aunque no sean números). Por ejemplo, el ranking de una crítica cinematográfica; la calificación de un producto,  $\mathcal{Q} = \{\text{malo}, \text{regular}, \text{bueno}\}$ .

En la práctica, la diferencia entre variables continuas y discretas no es muy precisa. Muchas veces, variables continuas son definidas como discretas, como por ejemplo, la edad en años o la talla en centímetros. Es importante tener presente el tipo de variable que se está estudiando, porque las herramientas utilizadas son diferentes <sup>3</sup>.

## 2.4 Distribución de frecuencias y sus representaciones gráficas

### 2.4.1 Caso de una variable discreta

Considerando los números de hermanos del ejemplo 2.1 (Tabla 2.1), notemos que algunos datos están repetidos. Es así que hay 5 alumnos que tienen 2 hermanos. Este número 5 es la **frecuencia** del valor 2 de la variable *Número de hermanos*. El conjunto de las frecuencias asociadas a cada valor se llama **distribución de frecuencias**, función que relaciona cada valor distinto que puede tomar la variable con el número

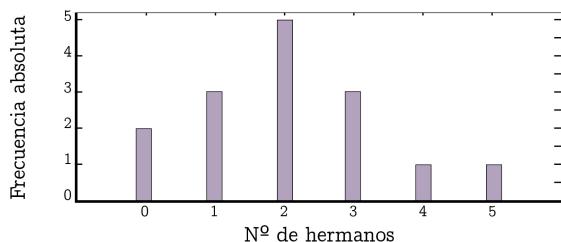
<sup>3</sup>Veremos que podemos transformar una variable cuantitativa para que sea nominal o ordinal. No se puede hacer lo contrario, salvo para codificar una variable con el objeto de facilitar su gestión, en particular para la digitación de los datos.

de veces que aparece este valor (Tabla 2.4), también llamado **frecuencia absoluta** del valor. Al pasar de los 15 valores de número de hermanos a las frecuencias, perdemos información, por ejemplo, ¿quién es el alumno que tiene 5 hermanos? o ¿quiénes son los alumnos que no tienen hermanos?, pero ganamos en interpretación: Reducimos la cantidad de valores a analizar para conservar lo esencial; por ejemplo, que hay 2 alumnos sin hermanos, que hay un alumno con 5 hermanos, o que el número de hermanos más frecuente es 2. Con la distribución de frecuencias vemos mejor cómo se distribuyen los 15 valores entre los 6 valores diferentes que puede tomar la variable. Esta distribución de frecuencias puede representarse también en un gráfico llamado **diagrama de barras** que muestra visualmente la relación entre cada valor tomado por la variable y su frecuencia (Figura 2.1). Los distintos valores tomados por la variable están en el eje de las abscisas y en cada valor se dibuja una barra de altura igual a la frecuencia correspondiente. La altura de la barra aparece, entonces, en la ordenada.

TABLA 2.4. Frecuencias absolutas

Nº Hermanos	0	1	2	3	4	5
Frecuencia absoluta	2	3	5	3	1	1

FIGURA 2.1. Diagrama de barras Nº hermanos



Si tenemos dos grupos de datos, uno con 15 individuos y otro con 100, será difícil comparar las dos distribuciones de frecuencias a partir de las frecuencias absolutas. Es preferible usar en este caso las **frecuencias relativas**, que son las frecuencias expresadas como proporciones de la cantidad total de observaciones (15 alumnos, en el ejemplo anterior), o como porcentajes (Tabla 2.5 y gráficos de la Figura 2.2). Se observa que los gráficos tienen el mismo aspecto, pero cambian los valores en la ordenada. Las frecuencias relativas suman 1 y los porcentajes suman 100. Si queremos saber cuántos alumnos tienen a lo más 1 hermano, tenemos que sumar las frecuencias absolutas para 0 y 1 hermanos:  $2+3 = 5$ . Las frecuencias absolutas acumuladas se

obtienen sumando las frecuencias absolutas que se encuentren contenidas hasta el valor considerado. Las frecuencias relativas acumuladas y los porcentajes acumulados se obtienen de manera similar (Tabla 2.6) con los gráficos asociados en la Figura 2.3.

Así podemos responder rápidamente a una de las preguntas iniciales, ¿la mayoría de los alumnos tienen a lo más 2 hermanos? En la Tabla 2.6 encontramos que 66.7% de los alumnos tienen a lo más 2 hermanos.

TABLA 2.5. Frecuencias relativas y porcentajes

Nº Hermanos	0	1	2	3	4	5
Frecuencia relativa	0,133	0,200	0,333	0,200	0,067	0,067
Porcentaje	13,3 %	20,0 %	33,3 %	20,0 %	6,7 %	6,7 %

FIGURA 2.2. Diagramas de barras

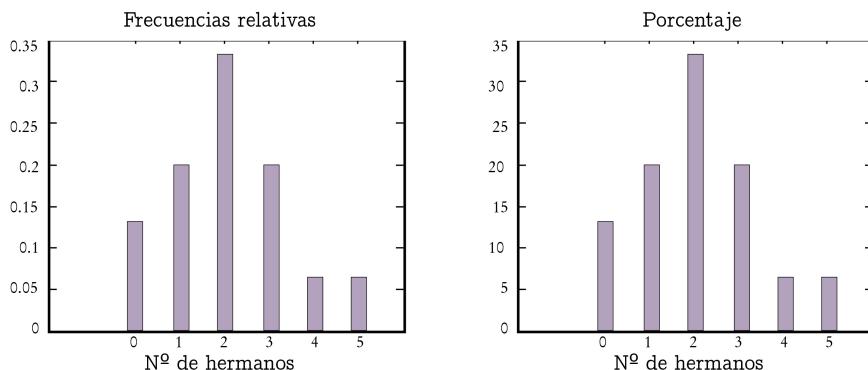
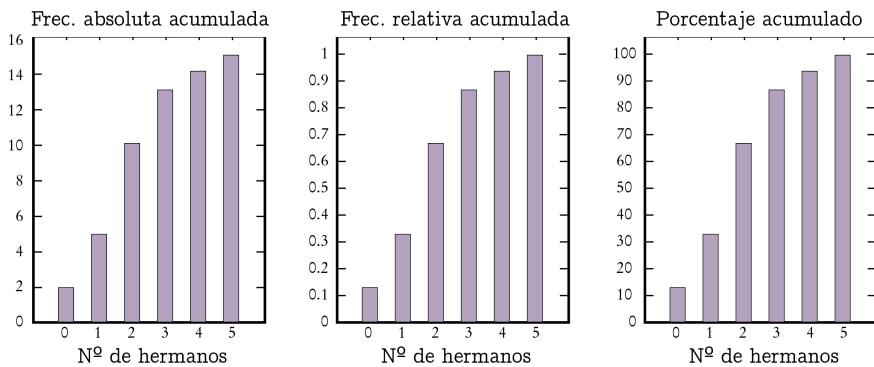


TABLA 2.6. Frecuencias relativas acumuladas

Nº Hermanos	0	1	2	3	4	5
Frecuencia absoluta acumulada	2	5	10	13	14	15
Frecuencia relativa acumulada	0,133	0,333	0,667	0,867	0,933	1,000
Porcentaje acumulado	13,3 %	33,3 %	66,7 %	86,7 %	93,3 %	100,0 %

FIGURA 2.3. Diagramas de barras frecuencias acumuladas



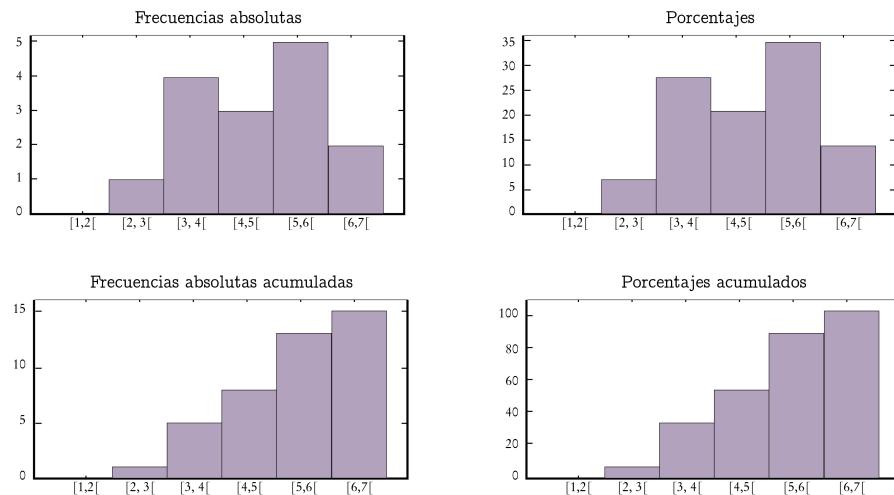
#### 2.4.2 Caso de una variable continua

Si consideramos ahora las notas de la primera prueba de matemáticas, observamos que no hay en general muchas notas repetidas. En este caso, las frecuencias definidas para cada valor distinto de la variable no son de mucho interés. Una manera de calcular frecuencias, como lo hicimos anteriormente con el número de hermanos, y que tengan una interpretación interesante, consiste en formar intervalos de notas: Por ejemplo [1,2[, [2,3[, ..., [6,7] y contar cuántos alumnos tienen notas de matemáticas menores que 2, entre 2 y 3, etc. De esta manera, obtenemos las frecuencias absolutas por intervalos en la Tabla 2.7. El intervalo [2,3[ contiene las notas mayores o iguales a 2 y menores que 3, etc. Observe que la nota 7,0 se cuenta en el intervalo [6,7]. Se deducen las frecuencias relativas no acumuladas y acumuladas y los porcentajes asociados (Tabla 2.7 y Figura 2.4). Se puede ver, por ejemplo, que 5 alumnos no alcanzan el 4,0 y que la mayoría tienen su nota entre 3,0 y 3,9 o entre 5,0 y 5,9.

TABLA 2.7. Frecuencias de las notas de la primera prueba de matemáticas

Nota	[1,2[	[2,3[	[3,4[	[4,5[	[5,6[	[6,7]
Frecuencia absoluta	0	1	4	3	5	2
Frecuencia relativa	0,0000	0,0667	0,2667	0,2000	0,3333	0,1333
Porcentaje	0,00 %	6,67 %	26,67 %	20,00 %	33,33 %	13,33 %
Frecuencia absoluta acumulada	0	1	5	8	13	15
Frecuencia relativa acumulada	0,0000	0,0667	0,3333	0,5333	0,8667	1,0000
Porcentaje acumulado	0,00 %	6,67 %	33,33 %	53,33 %	86,67 %	100,00 %

FIGURA 2.4. Histogramas de las notas de la primera prueba de matemáticas



Es importante notar que los intervalos de notas no pueden tener intersección, porque en ese caso estaríamos contando a algunos alumnos en más de un intervalo. Además, los intervalos tienen que cubrir todo el recorrido de valores, o sea **una nota cualquiera entre 1 y 7 tiene que poder clasificarse en un intervalo y sólo uno**. Cuando los diagramas se construyen con intervalos se habla de **histograma** y las barras se tocan. Observe que las barras no se tocan en el caso de un diagrama de barras, para mostrar que los valores que toma la variable son discretos.

#### Notas:

- El número de intervalos y sus límites son arbitrarios, pero deben elegirse con cuidado, para conseguir el efecto deseado, que es dar una visión rápida y sintética de la distribución de los datos.
- Es interesante observar que la construcción de intervalos tiene el efecto de transformar la *nota*, que es una variable “continua”, en una variable “discreta”.

Ahora consideraremos los resultados SIMCE (Ejemplo 2.3), cuya tabla de datos es demasiado grande para verla: 2389 colegios, lo que luce evidente el interés de reducir la información a lo esencial para poder mostrar su contenido e interpretarlo. Para definir intervalos de puntajes SIMCE para la prueba de matemática, tenemos que considerar el puntaje mínimo (158) y el máximo (380) y definir intervalos que cubren todos los puntajes entre 158 y 380 y que no tengan intersección. Podemos tomar 12 intervalos de anchos iguales desde el puntaje 155 hasta 395. Procedemos, entonces, como lo hicimos anteriormente con las notas de la primera prueba, pero como la tabla

es grande, más vale hacerlo con un computador; por ejemplo, con una planilla Excel. Se muestran las frecuencias absolutas, los porcentajes y los porcentajes acumulados obtenidos en la Tabla 2.8 y los gráficos de la Figura 2.5. Esta vez, se indicó el centro de los intervalos como abscisa de los gráficos.

TABLA 2.8. Frecuencias SIMCE colegio 2006 2do medio

Intervalo	Frecuencia absoluta	Porcentaje	Porcentaje acumulado
[155, 175[	14	0,586	0,586
[175, 195[	143	5,986	6,572
[195, 215[	358	14,985	21,557
[215, 235[	396	16,576	38,133
[235, 255[	358	14,985	53,118
[255, 275[	279	11,678	64,796
[275, 295[	240	10,046	74,842
[295, 315[	228	9,544	84,386
[315, 335[	219	9,167	93,553
[335, 355[	129	5,400	98,953
[355, 375[	24	1,005	99,958
[375, 395[	1	0,042	100,000
Total	2389	100	100

Si cambiamos el número de intervalos, la distribución puede tomar un aspecto diferente (Figura 2.6). ¿Qué pasa cuando se cambia el número de intervalos? Se puede también construir el diagrama de barras o el histograma horizontalmente (Figura 2.7).

Interprete la Tabla 2.9 y los gráficos 2.8 relativos a la PSU matemática del 2005.

Observe que la forma de la distribución de frecuencias no acumuladas del puntaje PSU es diferente de la del puntaje SIMCE (Figura 2.5). La primera se ve bastante más simétrica.

*Se dice que una distribución es simétrica si existe un valor V tal que el lado derecho del gráfico, a partir de V, es la imagen por un espejo del lado izquierdo (Figura 2.9).*

En las Figuras 2.10 se representan las frecuencias absolutas de las variables *conducta* y *género*. En la abscisa no aparecen números, sino los nombres de las categorías de la variables. En el caso de la variable *conducta* tiene sentido ordenar las categorías

FIGURA 2.5. Histogramas del promedio SIMCE 2do medio 2006 por colegio

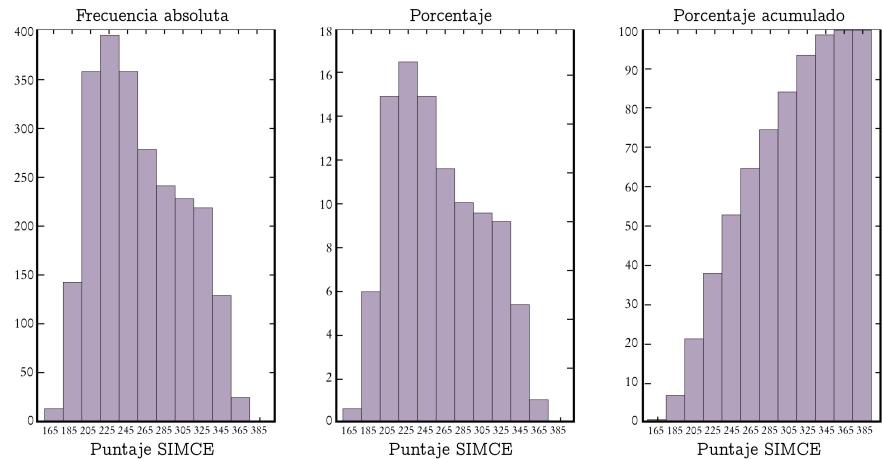
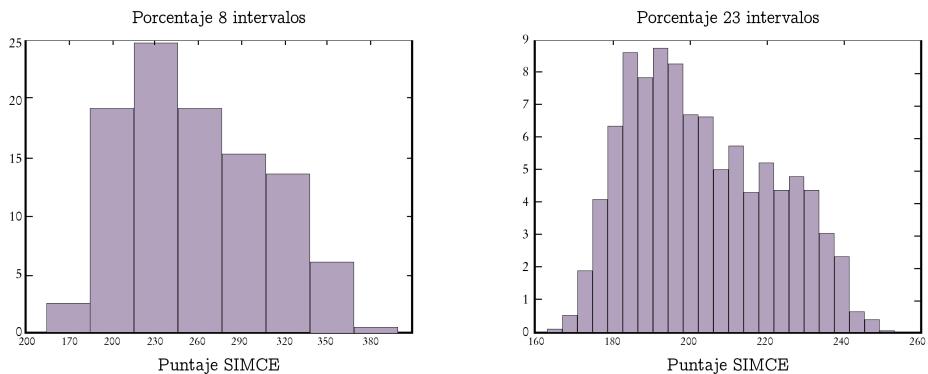


FIGURA 2.6. Histogramas del SIMCE con 8 y 23 intervalos



desde “Mala” a “Buena”, pero en el caso del género, el orden no tiene importancia. Podríamos mostrar también los diagramas de los porcentajes, pero ya sabemos que las formas son las mismas que las de los diagramas de las frecuencias absolutas o relativas.

#### 2.4.3 Diagrama de torta

Podemos mostrar los porcentajes como la división de una torta cuyos pedazos son proporcionales a los porcentajes y de manera que agotan la torta entera. O sea, un pedazo correspondiente a 20 % tiene un ángulo igual a  $20 \times 360 / 100 = 72$  grados. Son los **diagramas circulares**, llamados también **diagramas de torta**.

FIGURA 2.7. Histogramas SIMCE horizontal y vertical

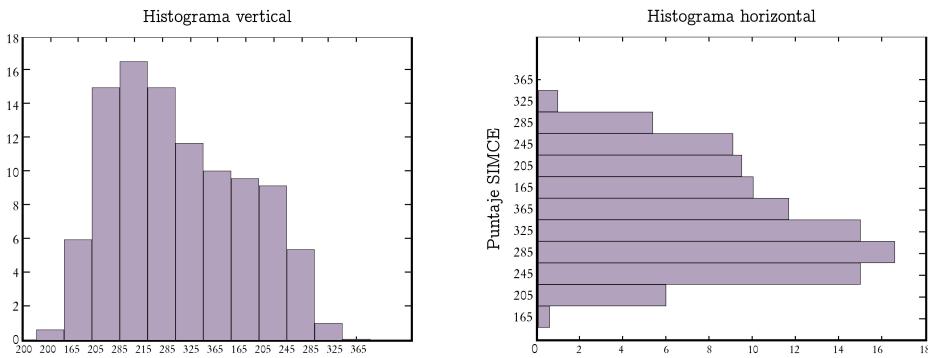


TABLA 2.9. Frecuencias PSU matemática 2005

Intervalo	Frecuencia absoluta	Porcentaje	Porcentaje acumulado
[0, 200[	926	0,545	0,545
[200, 250[	1248	0,734	1,279
[250, 300[	3270	1,924	3,203
[300, 350[	6503	3,825	7,028
[350, 400[	18736	11,022	18,050
[400, 450[	22971	13,513	31,563
[450, 500[	31434	18,491	50,054
[500, 550[	29444	17,321	67,375
[550, 600[	24723	14,543	81,918
[600, 650[	15489	9,111	91,029
[650, 700[	9048	5,323	96,352
[700, 750[	4342	2,554	98,906
[750, 800[	1522	0,895	99,801
[800, 850[	338	0,199	100,000

Pueden ver los gráficos de los porcentajes de alumnos por nivel de conducta o por género en la Figura 2.11. Estos gráficos dan una mejor idea de las frecuencias relativas que los diagramas de barra.

FIGURA 2.8. Histogramas de la PSU matemática 2005

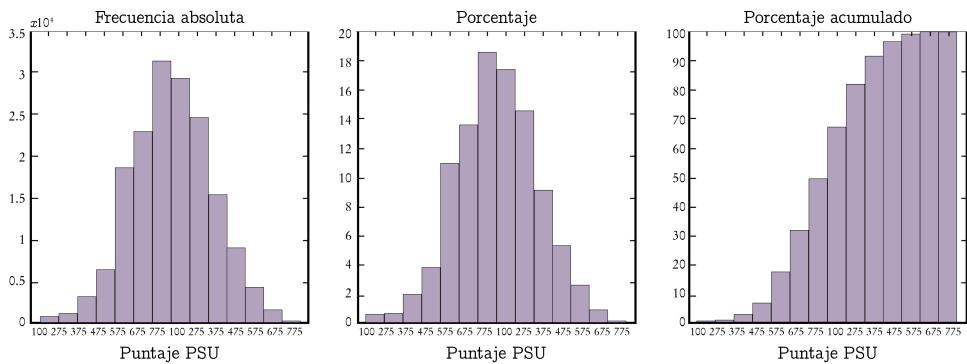
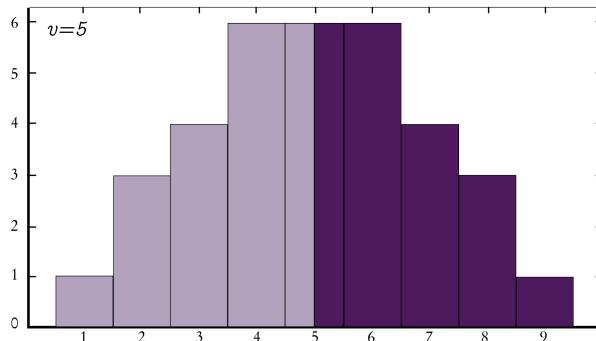


FIGURA 2.9. Histograma de una distribución simétrica



#### 2.4.4 Gráfico de tallo y hojas

Cuando el número de valores es pequeño, se puede visualizar la distribución de todos los datos usando un **gráfico de tallo y hojas**<sup>4</sup>. Tomando las 15 notas de la primera prueba ordenadas de menor a mayor:

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

Calculamos, entonces, la parte entera de cada una:

2 3 3 3 4 4 4 5 5 5 5 6 6,

y la diferencia entre la nota y su parte entera:

0,4 0,2 0,4 0,8 0,9 0,7 0,8 0,9 0,2 0,4 0,8 0,8 0,8 0,0 0,8.

<sup>4</sup>El gráfico de tallo y hojas se incluye aquí sólo porque se suele enseñar en la enseñanza media, aunque no es parte de los Contenidos Mínimos Obligatorios del MINEDUC. Este gráfico se usa muy poco.

FIGURA 2.10. Diagramas de barras de la conducta y del género

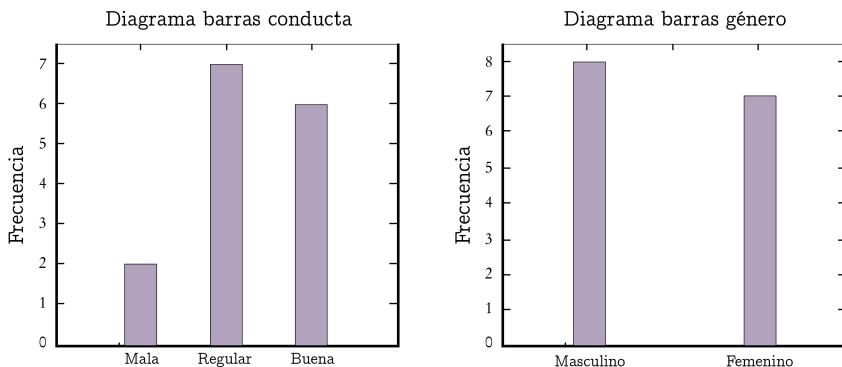
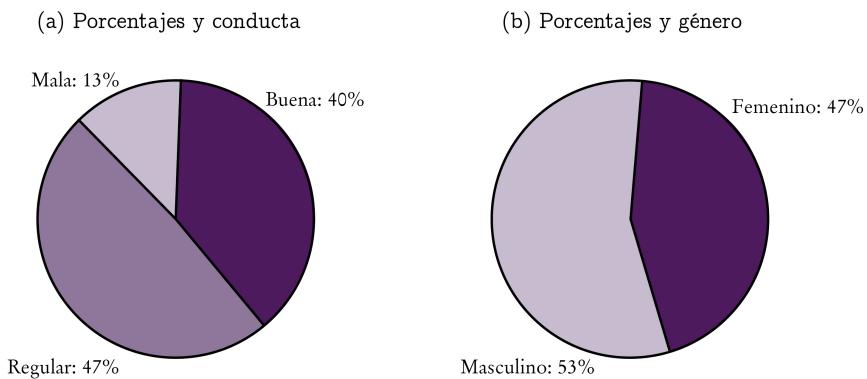
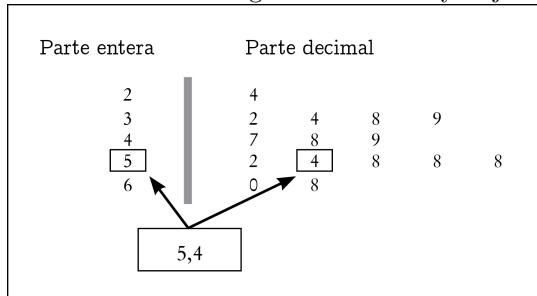


FIGURA 2.11. Diagramas circulares de la conducta y del género



Separamos las notas en paquetes, de tal manera que en un mismo paquete las notas tengan la misma parte entera. En un gráfico trazamos un recta vertical y ponemos a la izquierda las partes enteras (2, 3, 4, 5 y 6). Para las notas con parte entera igual a 2, en la misma altura que el dos, ponemos a la derecha, el valor 4, que corresponde a la parte decimal de las diferencias entre la nota y su parte entera:  $2,4 - 2 = 0,4$ . Para el paquete siguiente, parte entera igual a 3, se pone horizontalmente a la derecha del valor 3, los números 2, 4, 8 y 9, que son las decimales de las diferencias entre 3,2, 3,4, 3,8 y 3,9 y sus partes enteras 3, etc. Las partes enteras son los tallos y las partes decimales son las hojas del árbol (Figura 2.12). La acumulación de los datos da la forma de la distribución de frecuencias y los números permiten reconstituir los datos.

FIGURA 2.12. Un gráfico de tallos y hojas



## 2.5 Resúmenes de la distribución de frecuencias

### 2.5.1 Medidas de posición central

¿Es posible resumir los datos en forma aún más sintética que con una distribución de frecuencias? Si la variable es continua o discreta, lo más usual es hacerlo con un promedio. Todos hemos calculado promedios: el promedio de las notas en el colegio, el promedio de lo que gastamos al mes en movilización, en el supermercado, etc. Este número resulta de sumar todos los valores tomados por la variable y dividirla por el número de valores sumados. Pero, ¿qué nos dice el promedio? Veamos dos ejemplos:

- La nota promedio del curso del profesor Cabello en la sexta prueba de matemáticas es 5,7, ¿qué quiere decir?: “todos los estudiantes se sacaron un 5,7”. ¿Que la mayoría de los estudiantes obtuvo un 5,7? Sin embargo, nadie en el curso se sacó 5,7, ¿es esto posible?
- El promedio del número de hermanos es 2,07. Pero, ¿cómo puede un número con decimales informarnos sobre una variable que en la práctica sólo toma valores enteros? Todos tenemos amigos con 2 o bien 3 hermanos, pero nadie con 2,07 hermanos. En general, el promedio (u otros estadísticos) de una variable discreta no pertenece al conjunto de valores posibles de la variable misma.

¡Ya no es tan claro lo que es un promedio!

El promedio puede o no coincidir con uno de los valores utilizados en su cálculo. Se dice que es el centro físico de los datos. El concepto de objeto promedio o persona promedio está ilustrado en la figura adjunta, donde los objetos del platillo izquierdo son todos iguales al objeto promedio y los objetos del platillo derecho son los objetos de la población (6, 4, 9, 8, 6, 2, 10, 3). El promedio corresponde a un punto de equilibrio donde los valores mayores balancean a los menores.

Históricamente, un astrónomo y matemático belga Adolphe Quetelet (1796-1874) realizó los primeros intentos de aplicar la estadística a las Ciencias Sociales. Una de sus contribuciones fue el concepto de *persona promedio*, “persona cuya acción e ideas corresponde al resultado promedio obtenido sobre la sociedad entera”. Otra de sus contribuciones, que es menos conocida, es el Índice de Masa Corporal (IMC), el popular índice que mide la asociación entre el peso y la talla de un individuo.

Veamos como se escribe formalmente el promedio aritmético o media aritmética. Si tenemos  $n$  valores  $x_1, x_2, \dots, x_n$ , su promedio aritmético es:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Podemos decir que el promedio está bien definido y es fácil de calcular. Además, se presta bien a operaciones algebraicas y da cuenta de todos los valores.

El promedio indica el nivel o la posición de los datos. Por ejemplo, la nota de matemática de la primera prueba de la Tabla 2.1 tiene un promedio o media de 4,8. Podríamos argumentar, entonces, que 4,8 es una nota *típica* de la primera prueba del curso.

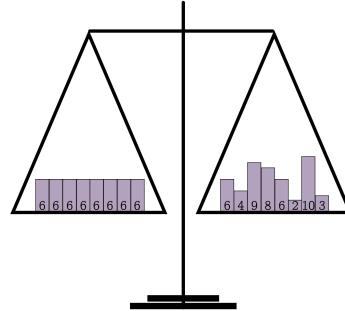
Una propiedad interesante del promedio aritmético es su **linealidad**, es decir, si se hace un cambio de origen de todos los valores, el promedio se traslada de la misma manera; si se hace un cambio de escala de los valores, el promedio tendrá la misma transformación. Si  $y_i = ax_i + b$ , para  $i = 1, 2, \dots, n$ , entonces el promedio de los  $y_i$  es:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = a\bar{x} + b.$$

La fácil demostración se deja al lector. En el Ejercicio 7 se muestra un ejemplo clásico.

Vemos que podemos calcular el promedio aritmético del número de hermanos a partir de las frecuencias relativas. En efecto, los datos son:

0 2 2 1 3 3 1 0 2 4 2 5 1 3 2



El promedio aritmético es:

$$\frac{0 + 2 + 2 + 1 + 3 + 3 + 1 + 0 + 2 + 4 + 2 + 5 + 1 + 3 + 2}{15} = 2,07$$

Si ordenamos los 15 datos:

0 0 1 1 1 2 2 2 2 3 3 3 3 4 5

Podemos calcular el promedio:

$$\frac{2 \times 0 + 3 \times 1 + 5 \times 2 + 3 \times 3 + 1 \times 4 + 1 \times 5}{15} = 2,07$$

$$\frac{2}{15} \times 0 + \frac{3}{15} \times 1 + \frac{5}{15} \times 2 + \frac{3}{15} \times 3 + \frac{1}{15} \times 4 + \frac{1}{15} \times 5 = 2,07$$

Consideremos ahora los alumnos de un curso de 1º medio de  $n$  alumnos. Si hay  $n_1$  niños y  $n_2$  niñas ( $n = n_1 + n_2$ ), la media aritmética  $\bar{x}$  de las notas de matemáticas de los  $n$  alumnos es igual a la media ponderada de las medias aritméticas de las notas de los niños y de las niñas  $\bar{x}_1$  y  $\bar{x}_2$ :

$$\bar{x} = \frac{n_1}{n} \bar{x}_1 + \frac{n_2}{n} \bar{x}_2.$$

### 2.5.2 Medias generalizadas

En ciertas ocasiones, puede ser útil aplicar una función  $f$  a los valores de  $x$  antes de calcular el promedio, y luego aplicar la transformación inversa al resultado. Se obtiene así la **media generalizada** de  $x$  según la transformación  $f$ :

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

Cuando  $f$  es la función identidad  $f(x) = x$ , el resultado se reduce al promedio aritmético que ya hemos visto. Otros casos particulares interesantes son

- $f(x) = x^2$ , que entrega la media cuadrática de  $x$ ;
- $f(x) = 1/x$ , que entrega la media armónica de  $x$ ;
- $f(x) = \log(x)$ , que entrega la media geométrica de  $x$ .

En efecto, en ciertas situaciones los valores de la distribución no son de naturaleza propiamente aditiva, como, por ejemplo, en el caso del índice de precios o la tasa de interés de un crédito hipotecario. Cuando se desea obtener promedios para este tipo de valores, que representan la evolución de una característica respecto a una situación anterior, es preferible utilizar la **media geométrica** como medida de posición central más representativa. La **media geométrica** de los datos  $x_1, \dots, x_n$  que corresponde a la transformación logarítmica de los datos:

$$m_g = \sqrt[n]{x_1 x_2 \dots x_n}.$$

Cabe notar que la media geométrica es más sensible que el promedio aritmético a los pequeños valores.

Consideramos ahora que usted recorre todos los días durante una semana la distancia  $\mathcal{D}$  entre Santiago y Valparaíso en su vehículo. Su velocidad no es la misma todo los días: el primer día es de 70 km/h, el segundo de 65 km/h, etc. Quiere saber cuál fue su velocidad promedio de la semana. El promedio aritmético o la media geométrica no son buenas medidas en este caso. Es preferible utilizar la **media armónica**, que corresponde a la transformación  $1/x$ :

$$m_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}.$$

En efecto la suma de las velocidades no tiene sentido. Lo que se puede sumar es los tiempos tomados para recorrer esta distancia, que correspondería al tiempo tomado para recorrer una distancia igual a  $7\mathcal{D}$ .

Si las velocidades diarias fueron  $v_1, v_2, \dots, v_7$ , entonces, los tiempos utilizados fueron:  $t_i = \frac{\mathcal{D}}{v_i}$ ,  $i = 1, 2, \dots, 7$ . Luego, el tiempo total utilizado por los 7 viajes fue:

$$T = \sum_{i=1}^7 t_i = \sum_{i=1}^7 \frac{\mathcal{D}}{v_i}$$

y la velocidad promedio es:

$$V = \frac{7\mathcal{D}}{T} = \frac{7\mathcal{D}}{\mathcal{D} \sum_{i=1}^7 \frac{1}{v_i}} = \frac{7}{\sum_{i=1}^7 \frac{1}{v_i}}.$$

Si las velocidades diarias fueron: 70, 65, 58, 72, 75, 64 y 68, la media armónica es igual a: 67,0082. El promedio aritmético, que vale 67,4286 y la media geométrica, que vale 67,2207, son mayores que la media armónica. Ambos estarían sobre-evaluando la media armónica, que es el promedio correcto aquí (ver ejercicios 13 y 14).

Un resultado interesante:

$$m_h \leq m_g \leq \bar{x}$$

No se demuestra aquí.

El estadístico escocés Georges U. Yule estableció condiciones que debería cumplir una buena medida que representa la posición y la dispersión de una distribución empírica en su libro “An Introduction to the theory of statistics”:

- tener objetividad;
- depender de todas las observaciones;
- tener un significado concreto;
- ser fácil de calcular;
- ser poco sensible a fluctuaciones debidas al muestreo.

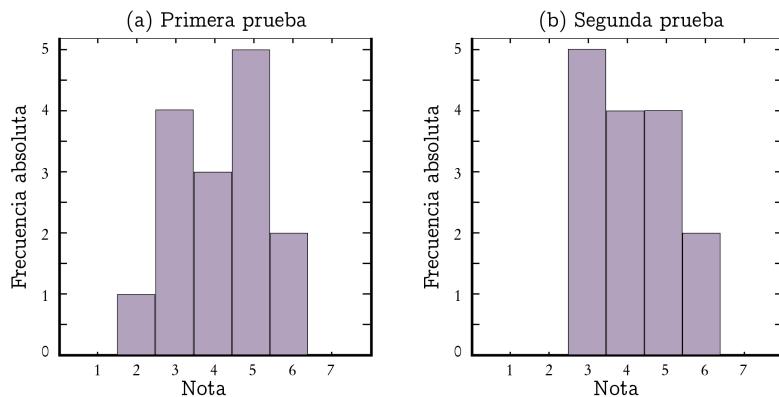
En la práctica, es imposible encontrar medidas que cumplan todas las condiciones. La elección de una medida debe considerar el estudio que la utilizará. En ciertos casos es la facilidad de comunicar los resultados, que es importante, en otros, es la necesidad

de cálculos simples, y en encuestas por muestreo, será más bien la sensibilidad a las fluctuaciones del muestreo (Capítulo 3).

### 2.5.3 Medidas de dispersión

Calculemos ahora el promedio de las notas de la segunda prueba del curso del profesor Cabello. Encontramos que es igual al de la primera prueba, ¿esto significa que todas las notas de la segunda prueba son iguales a las de la primera? Claramente no, entonces ¿son iguales las dos distribuciones de frecuencias? No necesariamente (Figura 2.13).

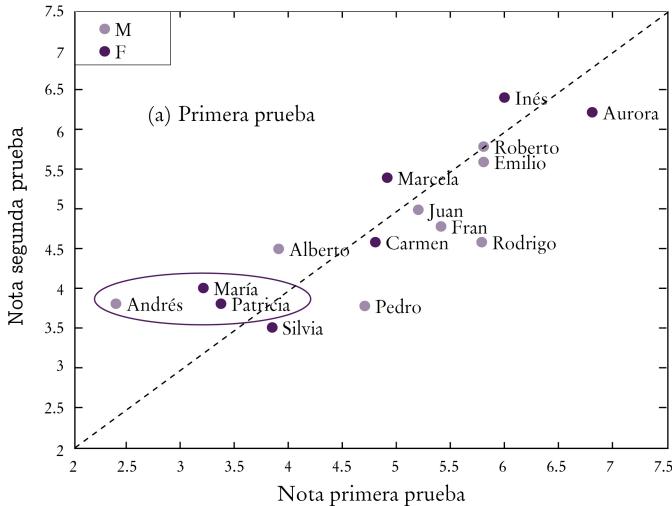
FIGURA 2.13. Notas dos primeras pruebas



Podemos comparar las notas de las dos pruebas mediante un **gráfico de dispersión** (Figura 2.14), donde se representa a cada alumno como un punto del plano. La nota en la prueba 1 corresponde a la abscisa y la nota en la prueba 2, a la ordenada. Los alumnos con notas parecidas en ambas pruebas quedan cerca de la diagonal. Los alumnos que tienen menor nota en la segunda prueba se encuentran por debajo de la diagonal y los alumnos que tienen mayor nota en la segunda prueba, por arriba de la diagonal. Los alumnos Andrés, María y Patricia tienen una nota bastante mayor en la segunda prueba que en la primera. Las notas de ambas pruebas fluctúan alrededor del mismo valor, el promedio 4.8, pero las notas de la Prueba 1 se dispersan más que las de la Prueba 2. En este gráfico se puede estudiar también las diferencias entre género.

El concepto de **dispersión** es una característica importante que complementa el concepto de promedio para describir la distribución de frecuencias. Las medidas de posición central describen las observaciones “en general” o “en promedio”. Las medidas de dispersión nos informan hasta qué punto estas observaciones son cercanas o alejadas de su “media”.

FIGURA 2.14. Notas de las dos primeras pruebas



Las dos primeras maneras naturales de definir la dispersión de una distribución son:

- El **recorrido**: el intervalo comprendido entre el valor mínimo y el valor máximo;
- La **varianza**: cuantifica la variabilidad de un conjunto de datos midiendo su dispersión cuadrática  $s_x^2$  alrededor del promedio  $\bar{x}$ :

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

La unidad de medición de la varianza es el cuadrado de la unidad de medición de la variable. Para tener la misma unidad que la variable, basta tomar la **desviación estándar** que es  $s_x$  la raíz cuadrada de la varianza.

La varianza o la desviación estándar no tienen la propiedad de linealidad como la media. En efecto, si consideramos la transformación  $y_i = ax_i + b$ ,  $i = 1, 2, \dots, n$ , vemos que la media de los  $y_i$  es igual a  $\bar{y} = a\bar{x} + b$ , ¿qué pasa con la varianza y la desviación estándar?

$$\text{Var}(y) = s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (ax_i + b - a\bar{x} - b)^2 = \frac{a^2}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = a^2 \text{Var}(x). \quad (2.1)$$

Para la desviación estándar consideramos

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

Los recorridos de las primera y segunda pruebas son: 4,4 y 2,8, respectivamente, y las desviaciones estándares de las primera y segunda pruebas son 1,185 y 0,887 respectivamente. Estos números expresan bien lo que vimos gráficamente, a saber, que la primera prueba tiene una variabilidad más extensa que la segunda.

Una ventaja de la desviación estándar sobre el recorrido es que usa todos los valores mientras el recorrido usa solamente el mínimo y el máximo. No refleja lo que puede pasar entremedio. Además, es muy sensible a **valores atípicos** (muy extremos). Basta que un alumno tenga una nota muy baja o muy alta para cambiar el recorrido de las notas de todo el curso.

Una propiedad que hace coherente el par *promedio* y *desviación estándar* es la siguiente: El promedio aritmético es el valor  $b$  que minimiza la cantidad

$$CM = \frac{1}{n} \sum_{i=1}^n (x_i - b)^2. \quad (2.2)$$

Para demostrar este resultado basta derivar con respecto a  $b$  la expresión 2.2. El mínimo de la expresión es, entonces, la varianza. En resumen

- La suma de las diferencias entre los valores y la media aritmética es nula.
- La suma de los cuadrados de las diferencias entre los valores y la media aritmética, es menor que la suma de las diferencias entre los valores y cualquier otro valor distinta de la media aritmética.
- La media aritmética de una población compuesta de dos sub-poblaciones puede expresarse, de manera simple, como una media aritmética ponderada de las medias aritméticas de las dos sub-poblaciones.

La elección de la varianza o desviación estándar para medir la dispersión de una muestra es, por tanto, coherente con la elección del promedio aritmético.

#### 2.5.4 Otras medidas de posición y dispersión

Ordenando las notas de la primera prueba de menor a mayor, podemos decir que la nota 4,9 es una *nota intermedia* en el sentido que hay tantas notas inferiores como superiores a 4,9 (Figura 2.15). Este número se llama **mediana**. Es otra manera de definir el centro de una distribución de frecuencias.

Para calcular la mediana, se ordenan los datos de menor a mayor. Las posiciones de las observaciones después de ordenarlas se llaman **rangos**. La mediana depende, entonces, de si el número  $n$  de observaciones es par o impar:

- (1) Si  $n$  es impar, se calcula  $k = \frac{n+1}{2}$ . La mediana es el valor de la observación de rango  $k$ , que deja tantos valores por debajo como por encima de él.
- (2) Si  $n$  es par, se calcula  $k = \frac{n}{2}$ . Cualquier valor entre los datos de rangos  $k$  y  $k + 1$  deja tantos valores por debajo como por encima de él. Pero nosotros

daremos en forma convencional que la mediana es, en este caso, el valor de rango  $k + 1$ .<sup>5</sup>

En la Figura 2.15 tenemos las 15 notas de la primera prueba ordenadas de menor a mayor. La mediana es la nota del alumno de rango  $16/2=8$ . En la Figura 2.16 tenemos las 14 notas ordenadas de menor a mayor. La mediana es 5,1.

FIGURA 2.15. Notas de la primera prueba

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2,4	3,2	3,4	3,8	3,9	4,7	4,8	4,9	5,2	5,4	5,8	5,8	5,8	6,0	6,8
↑														
Mediana														
4,9														

FIGURA 2.16. Notas de otra prueba

1	2	3	4	5	6	7	8	9	10	11	12	13	14
2,4	3,2	3,4	3,8	3,9	4,7	4,8	5,1	5,2	5,4	5,8	5,8	5,8	6,0
↑													
Mediana													
5,1													

Como el promedio aritmético, la mediana conserva los cambios de origen y de escala. Se deja al lector verificar esta propiedad de linealidad. Generalmente, el promedio aritmético y la mediana son vecinos. La mediana tiene la ventaja de ser menos sensible a los valores extremos que el promedio, aunque no usa los valores propiamente tales. En efecto, supongamos que, en la primera prueba, el alumno de rango 5 tiene un 4,4 en vez de un 3,9. Calcule el promedio y la mediana de las nuevas notas. Constatará que el promedio cambia pero la mediana no. La mediana no es sensible a pequeños cambios. Se dice que la mediana es más *robusta* que el promedio.

De la misma manera que asociamos al promedio aritmético la desviación estándar como medida de dispersión, se asocia a la mediana  $M$  la medida  $DA$ , llamada **desviación absoluta**, igual al promedio de las diferencias absolutas de los valores con su mediana  $M$ :

$$DA = \frac{1}{n} \sum_{i=1}^n |x_i - M|.$$

<sup>5</sup>Algunos programas computacionales usan convenciones diferentes. La razón para preferir la convención propuesta aquí quedará clara más adelante, cuando hablemos de los cuantiles.

Compare la desviación absoluta con la desviación estándar de las notas de la prueba 6.

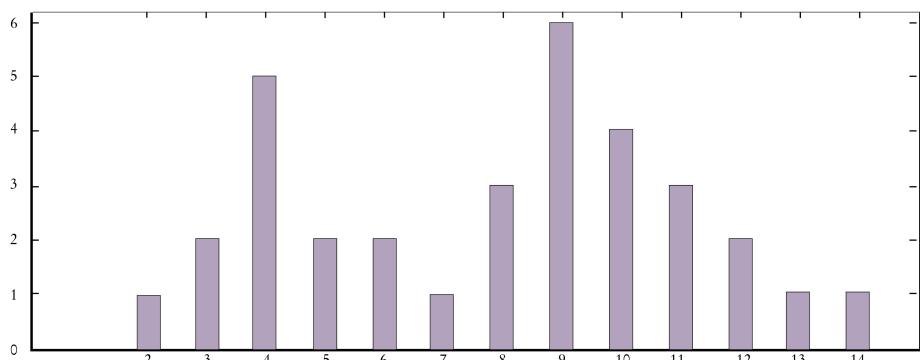
La mediana hace mínima la expresión :

$$\frac{1}{n} \sum_{i=1}^n |x_i - b|. \quad (2.3)$$

tal como el promedio aritmético hace mínima la desviación cuadrática (ver el ejercicio 15).

Ahora, consideremos el promedio del número de hermanos de los alumnos del profesor Cabello, que es 2,07. Se podría argumentar que nadie puede tener 2,07 hermanos, y que la posición central de la distribución habría que resumirla, en este caso, con un entero. Una medida de posición que consigue este efecto es la llamada **moda**, o valor más frecuente de la distribución, que en el caso de los alumnos de Cabello es 2. Las medidas de tendencia central - promedio, mediana y moda - pueden ser insuficientes en ciertos casos (Figura 2.17). En este gráfico, el promedio o la mediana no reflejan bien la distribución y los valores 4 y 9 tienen alta frecuencia. Se dice que la distribución es bimodal. Una distribución puede ocasionalmente tener más de una moda, lo que sugiere que hay dos o más *poblaciones* diferentes mezcladas.

FIGURA 2.17. Distribución bimodal



Un índice que expresa la relación entre posición y dispersión de la distribución es el **coeficiente de variación**, igual a la razón entre la desviación estándar y el promedio. Permite tener una proporción de la variabilidad con respecto a la magnitud de los valores tomados por la variable.

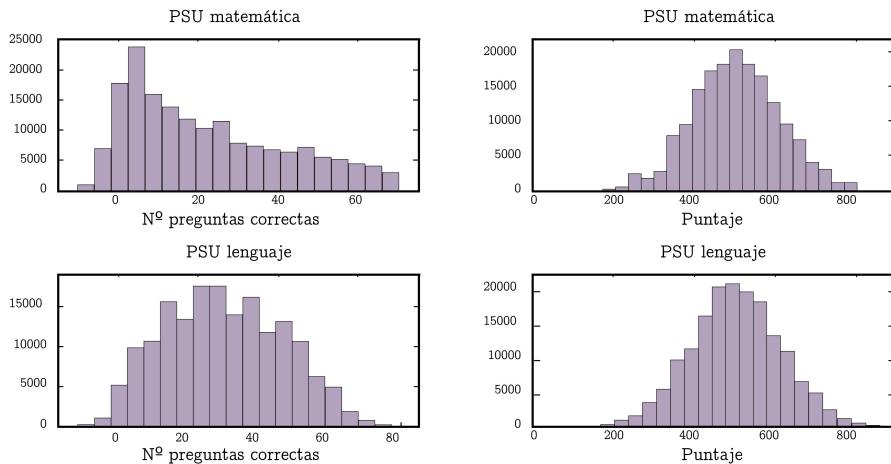
En resumen, hemos definido hasta ahora tres medidas (o estadígrafos) que indican la posición central de una distribución de frecuencias: el promedio, la mediana y la moda. Las tres medidas son interesantes y cada una permite decir algo distinto sobre la distribución. Las posiciones relativas de las tres medidas dependen de la forma de

la distribución. En caso de una distribución perfectamente simétrica y unimodal, el promedio aritmético, la mediana y la moda son iguales.

Las figuras 2.5 y 2.8 muestran que la distribución del puntaje de la PSU es simétrica y la del puntaje SIMCE no lo es. En el caso de la PSU esto sucede porque las autoridades transforman de oficio las notas de la PSU para que la distribución sea simétrica. La distribución del número de respuestas correctas en Matemática no es simétrica (Figura 2.18).

Veamos las estadísticas del SIMCE y de la PSU en la Tabla 2.10. Si la distribución está más cargada hacia el lado de los valores inferiores, como es el caso del número de respuestas correctas de la PSU matemática, la mediana resulta inferior al promedio. Es lo que observamos con la PSU matemática donde, el promedio es de 22,4 preguntas y la mediana 18,0 preguntas, lo que significa que la mitad de los alumnos contesta como máximo 18 preguntas correctamente (y no 22,4 preguntas, que es el promedio). La distribución de respuestas correctas para la PSU de lenguaje es más simétrica, y de hecho el promedio y la mediana son muy parecidas: 31,3 y 31,0 preguntas.

FIGURA 2.18. Distribuciones SIMCE y PSU



## 2.6 Diagrama de caja y el resumen de cinco números

Hemos visto que el promedio y la mediana, junto con sus correspondientes medidas de dispersión, dan una idea de la forma de la distribución, pero si nos interesáramos en saber cuántas preguntas contesta el 5% de los mejores alumnos, ni el promedio ni la mediana nos lo dirían. Incluso el histograma completo de toda la distribución podría ser insuficiente; esto depende del número de intervalos.

Pero podríamos encontrar el valor buscado, procediendo en forma análoga que para el cálculo de la mediana. Si tenemos  $n = 170.000$  alumnos que rindieron la PSU,

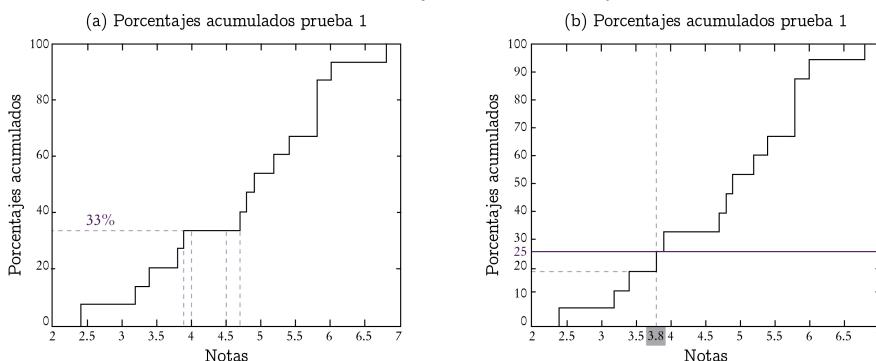
TABLA 2.10. Estadísticas SIMCE y PSU

	SIMCE		PSU matemática		PSU lenguaje	
	Matemática	Lenguaje	Preguntas correctas	Puntaje	Preguntas correctas	Puntaje
Promedio aritmético	257,1	258,3	22,4	500,0	31,3	500,0
Mediana	150,0	255,0	18,0	500,0	31,0	500,0
Moda	207,0	221,0	19,0	500,0	24,0	500,0
Desviación estándar	46,9	33,1	19,5	110,0	16,3	110,0
Desviación absoluta	39,4	27,9	16,1	86,7	13,7	87,6

el 5 % corresponde a 8.500 alumnos y el 95 % a  $170.000 - 8.500 = 161.500$  alumnos. Si ordenamos de menor a mayor los 170.000 valores, encontramos que el alumno de rango 161.500 contestó correctamente 60 preguntas. Se concluye que un 95 % de los alumnos contestó correctamente a lo más a 60 preguntas. Este valor de 60 preguntas se llama cuantil de orden 95 %.

Consideramos el gráfico de **función porcentajes acumulados** de las notas de la primera prueba (Figura 2.19 (a)). La función hace corresponder un porcentaje a cada nota. Para encontrar un cuantil necesitamos la **función inversa** de la función porcentajes acumulados. El problema es que no es biyectiva cuando se calcula a partir de datos empíricos. En efecto, vemos que las notas 4 y 4,5 tienen el mismo porcentaje acumulado (33 %) debido a que no hubo notas entre 4 y 4,5, ¿cómo construir una función inversa en este caso? Construir una función inversa significa atribuir una nota única a cada porcentaje acumulado. Ya vimos que para 33 %, cualquier nota entre 3,9 y 4,7 es posible.

FIGURA 2.19. Porcentajes acumulados y cuantiles



Usaremos, entonces, la siguiente convención:

Definición: Sean  $x = \{x_1, \dots, x_n\}$  un conjunto de datos **ordenados en orden creciente**. La función cuantil empírico de los datos es la función  $C$  tal que, para todo  $i = 1, \dots, n$ , vale  $x_i$  en el intervalo  $\left] \frac{i-1}{n}, \frac{i}{n} \right]$ . Vale decir, que  $\forall u \in \left] \frac{i-1}{n}, \frac{i}{n} \right], C(u) = x_i$ .

Hay que observar que si  $u = \frac{i-1}{n}$ , entonces,  $C(u) = x_{i-1}$ .

Por ejemplo, en el caso de las notas de la primera prueba, el cuantil 25% es 3,8, pero cualquier cuantil de 20% hasta 26% es 3,8 también. Pero el cuantil 19% es 3,4 y el de 27% es 3,9 (Figura 2.19 (b)).

Es fácil ver que la mediana también es un cuantil: el cuantil 50%. Se acostumbra usar los cuartiles 25%, 50% y 75%, llamados **primer cuartil**, **segundo cuartil** (o mediana) y **tercer cuartil**, respectivamente. Los denotaremos  $Q_1$ ,  $Q_2$  y  $Q_3$ . El espacio entre el primer y el tercer cuartil ( $Q_3 - Q_1$ ) se llama **intervalo intercuartílico**: es el intervalo que contiene la mitad central de los datos. Para las notas obtenemos los valores 3,8; 4,9 y 5,8, respectivamente, con un intervalo intercuartílico igual a 2 (Figura 2.20). Para obtener estos valores usted mismo, abra una planilla Excel y escriba en la primera línea, los rangos de 1 a 15 y en segunda línea los valores ordenados de las 15 notas. En la tercera línea escribe la función porcentaje acumulado de la siguiente manera: en la celda bajo de la nota mínima (2,4) escribe la formula: **=100\*A1/15** y cópiela debajo de las otras notas. Busque ahora entre qué porcentajes se encuentra el 25%. Aquí es entre los valores 3,4 y 3,8. Se toma entonces el valor 3,8, que corresponde a la cota superior del intervalo [3,4 y 3,8]. Se repite de manera similar para la mediana y el tercer cuartil.

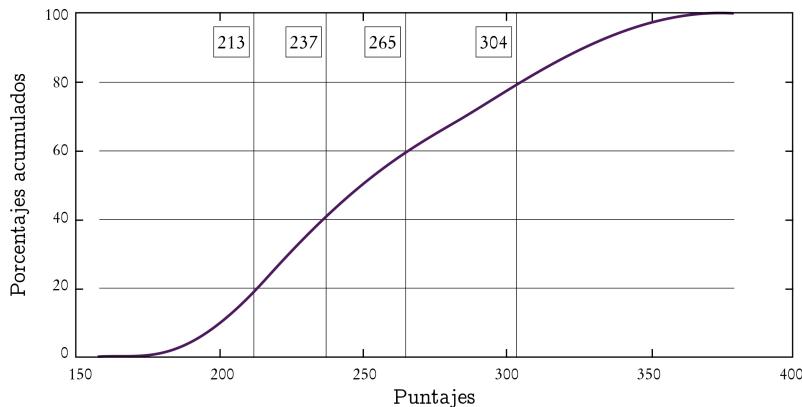
FIGURA 2.20. Cuartiles notas

Rangos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Valores	2,4	3,2	3,4	3,8	3,9	4,7	4,8	4,9	5,2	5,4	5,8	5,8	5,8	6,0	6,8
Porcentajes	6,7	13,3	20,0	26,7	33,3	40,0	46,7	53,3	60,0	66,7	73,3	80,0	86,7	93,3	100,0
				↑			↑				↑				
				1 <sup>er</sup> cuartil			2 <sup>do</sup> cuartil				3 <sup>er</sup> cuartil				
Cuartiles				3,8			4,9				5,8				

Se definen de la misma manera los quintiles (cada quintil corresponde a 20% del total), los deciles (cada decil corresponde a 10% del total) y los centiles (cada centil corresponde a 1% del total). En la figura 2.21, se muestran los quintiles de los resultados SIMCE 2006. Encontramos que 20% de los colegios tienen menos de 213 puntos, 40% menos de 237, 60% menos de 265 y 80% menos de 304. La distribución no es simétrica y está cargada a la izquierda: los puntajes altos se dispersan más.

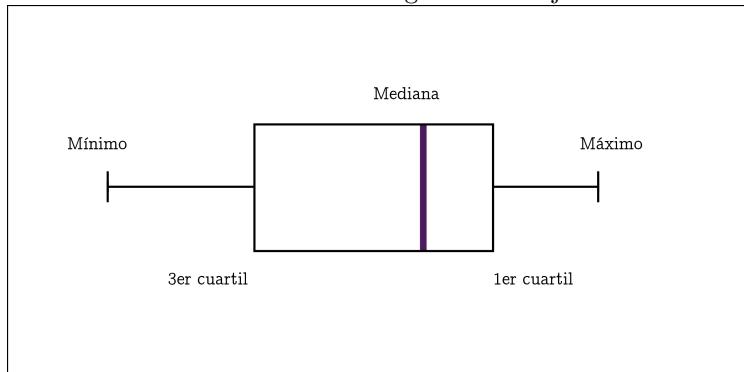
Un gráfico muy interesante es el llamado gráfico o diagrama de caja con bigotes o Box plot, o simplemente gráfico de caja que representa la distribución mediante el **resumen de cinco números: el mínimo, los tres cuartiles y el máximo**. En la Figura 2.22 se muestra un gráfico de caja que permite visualizar el recorrido y los cuartiles de la distribución, que son valores puestos en abscisa. En la caja hay 50%

FIGURA 2.21. Quintiles SIMCE 2006



de los valores más centrales, la línea roja que cruza la caja es la mediana. En la parte izquierda está el mínimo y en la derecha el máximo. Los segmentos que salen de la caja hacia el mínimo o al máximo se llaman *bigotes*.

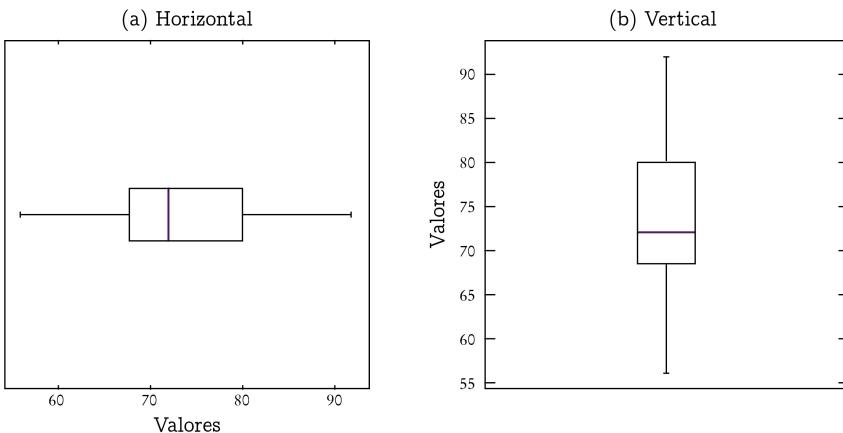
FIGURA 2.22. Un gráfico de caja



Sean los 31 datos siguientes:  $\{56, 56, 60, 60, 64, 64, 64, 68, 68, 68, 68, 68, 68, 72, 72, 72, 72, 72, 76, 76, 76, 80, 80, 80, 84, 84, 84, 88, 92\}$ . El valor mínimo es 56, el máximo es 92 y los tres cuartiles son 68, 72 y 80, respectivamente. El diagrama de caja se muestra en la Figura 2.23(a). El diagrama puede presentarse verticalmente también (Figura 2.23(b)).

La posición de la mediana en la caja refleja el grado de simetría de la distribución. Vemos que la mediana está más cercana al primer cuartil que al tercero; la distribución está más cargada del lado de los valores inferiores.

FIGURA 2.23. Diagrama de cajas



La gran ventaja del diagrama de caja es que permite comparar de manera simple varias distribuciones al mismo tiempo, como lo vemos en la Figura 2.24(a) donde aparece un diagrama de caja del resultado SIMCE por tipo de dependencia del colegio. Permite ver, en particular, que el intervalo intercuartílico del SIMCE de los colegios particulares pagados (PP) es más amplio que el de los colegios municipales (Mu). Interprete este gráfico.

Observemos que el primer cuartil para los colegios particulares pagados se distancia de la mediana mucho más que el tercer cuartil. Para los colegios municipales pasa lo contrario: el tercer cuartil se extiende más que el primero. Para entender lo que ocurre veamos el histograma de cada grupo de colegios (Figura 2.25) donde observamos que solamente para los colegios particulares subvencionados, la distribución muestra cierta simetría. En el gráfico (a) se observa que incluso si algunos colegios municipales alcanzan un alto puntaje SIMCE, éstos son pocos, y que muchos tienen valores más bien bajos. Ocurre lo contrario para los colegios particulares pagados. Los pocos colegios municipales con alto puntaje y que se alejan del resto pueden considerarse como **atípicos**. Los pocos colegios particulares pagados con bajo puntaje y que se alejan del resto son atípicos también. No hay una manera única y precisa que permita definir un valor atípico. Lo que está claro es que tiene que ver con la dispersión de los valores. Se puede usar como referencia la desviación estándar o (como lo hicimos aquí) el intervalo intercuartil.

Considerando los cuartiles  $Q_1$  y  $Q_3$ , se dice, en este caso, que un valor  $x$  es atípico,

- si  $x$  es un valor alto y  $x - Q_3 > \alpha(Q_3 - Q_1)$
- o si  $x$  es un valor bajo y  $Q_1 - x > \alpha(Q_3 - Q_1)$

donde  $\alpha$  es un valor por determinar. Generalmente, se toma un valor entre 1,5 y 3.

Los bigotes tienen, en este caso, un largo determinado por  $\alpha$  y el intervalo intercuartilico  $Q3 - Q1$ . Los valores atípicos son, entonces, representados por una cruz (Gráfico 2.24(b)). Este último gráfico hace mucho más evidente la brecha entre los colegios municipales y particulares pagados.

FIGURA 2.24. Gráficos de cajas SIMCE

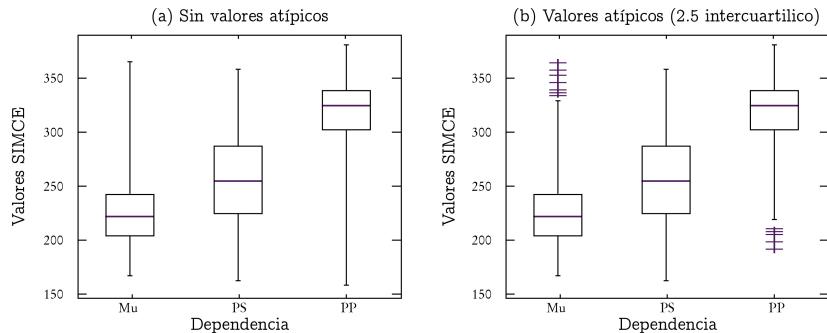
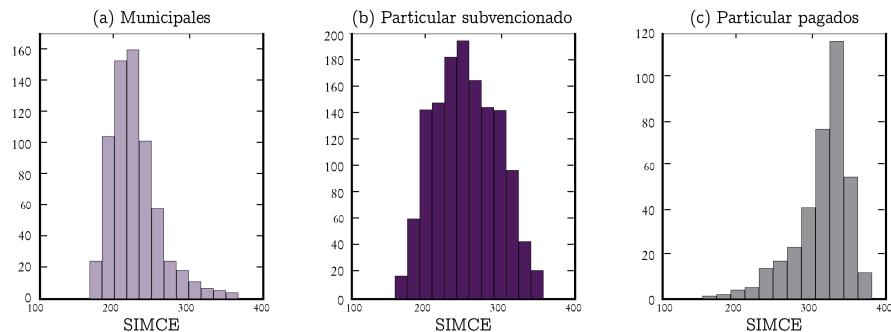


FIGURA 2.25. Histogramas SIMCE



## 2.7 Cómo visualizar dos o más variables al mismo tiempo

Vimos en el párrafo anterior, el gráfico de dispersión de las notas de las dos primeras pruebas del profesor Cabello (Figura 2.14). Eso nos permitió:

- tener una visualización de la *relación* entre dos variables cuantitativas,
- tener una configuración de grupos de individuos.

En el párrafo anterior acabamos de comparar las distribuciones del puntaje SIMCE, según la dependencia del colegio usando diagramas de cajas. Eso nos permitió:

- ver si la dependencia influye sobre la posición de la distribución del puntaje SIMCE,
- ver si la dependencia influye sobre la variabilidad del puntaje SIMCE.

El gráfico de dispersión (también llamado dispersiograma) permite relacionar dos variables cuantitativas y el diagrama de caja una variable cuantitativa con una variable nominal.

Para relacionar dos variables nominales pueden usarse las llamadas tablas cruzadas (o cruces de variables). En la Tabla 2.11 se encuentran las frecuencias absolutas del número de alumnos que rindieron la prueba SIMCE 2006, por tipo de dependencia y nivel socioeconómico (NSE). El nivel A es el nivel socioeconómico más bajo y E el más alto. Por ejemplo, encontramos 32918 alumnos de los colegios municipales de NSE muy bajo, 760 de colegios particulares subvencionados de NSE muy alto o ningún alumno de colegio particular pagado de NSE muy bajo. En la Tabla 2.12 se muestran las frecuencias relativas, lo que permitiría comparar las distribuciones cruzadas de la dependencia con el NSE de un año a otro, por ejemplo, o entre regiones, eliminando el efecto del número total de alumnos.

La Tabla 2.13 muestra los porcentajes de alumnos por NSE para cada tipo de dependencia y la Tabla 2.14 muestra los porcentajes de alumnos por dependencia para cada nivel socioeconómico. Por ejemplo, en la Tabla 2.14, las columnas de la tabla muestran cómo se distribuyen los alumnos entre los colegios municipales, particulares subvencionados y pagados de cada NSE. Estas dos tablas permiten comparar las distribuciones fijando una categoría. Interprete las dos tablas junto con los gráficos de la Figura 2.26.

TABLA 2.11. Frecuencias absolutas SIMCE según dependencia y NSE

	A	B	C	D	E	Total
Municipal	32918	50517	17106	2766	0	103307
Part. Subvencionado	12288	46785	41029	19945	760	120807
Part. Pagado	0	0	0	2045	15461	17506
Total	45206	97302	58135	24756	16221	241620

Finalmente en la Tabla 2.15 se muestran los promedios SIMCE por dependencia y NSE. La tabla cruza dos variables nominales (5 niveles para el NSE y 3 tipos de dependencia), lo que define 15 grupos de colegios. Se presenta el promedio SIMCE en cada grupo. Los gráficos asociados se encuentran en los gráficos de la Figura 2.27. Podemos observar que no se encuentra colegios municipales en la categoría NSE más alta, hay muy pocos colegios particulares subvencionadas en esta categoría alta y los colegios particulares pagados se encuentran solamente en las categorías más altas (D y E). Respecto de los promedios SIMCE vemos un “efecto” de la dependencia y del NSE. Cualquiera sea el tipo de dependencia, cuando aumenta el NSE, aumenta

TABLA 2.12. Porcentajes alumnos SIMCE según dependencia y NSE

	A	B	C	D	E	Total
Municipal	13,62	20,91	7,08	1,14	0,0	42,76
Part. Subvencionado	5,09	19,36	16,98	8,25	0,31	50,00
Part. Pagado	0	0	0	0,84	6,40	7,25
Total	18,71	40,27	24,06	10,25	6,71	100,00

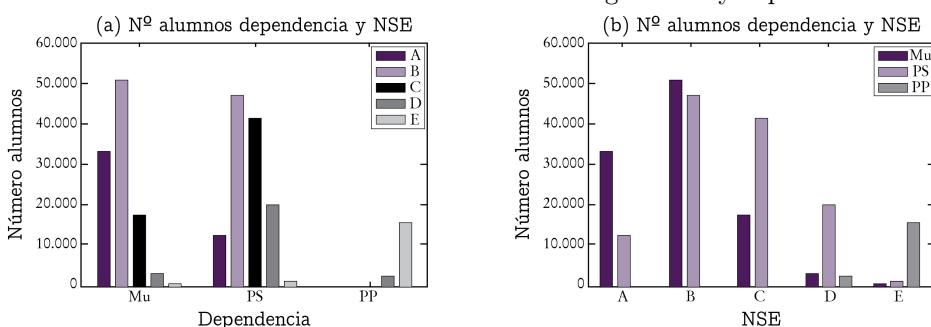
TABLA 2.13. Porcentajes filas

	A	B	C	D	E	Total
Municipal	31,86	48,90	16,56	2,68	0,00	100,00
Part. Subvencionado	10,17	38,73	33,96	16,51	0,63	100,00
Part. Pagado	0,00	0,00	0,00	11,68	88,32	100,00
Total	18,71	40,27	24,06	10,25	6,71	100,00

TABLA 2.14. Porcentajes columnas

	A	B	C	D	E	Total
Municipal	72,82	51,92	29,42	11,17	0,00	42,75
Part. Subvencionado	27,18	48,08	70,58	80,57	4,69	50,00
Part. Pagado	0,00	0,00	0,00	8,26	95,31	7,25
Total	100,00	100,00	100,00	100,00	100,00	100,00

FIGURA 2.26. Frecuencias alumnos SIMCE según NSE y dependencia

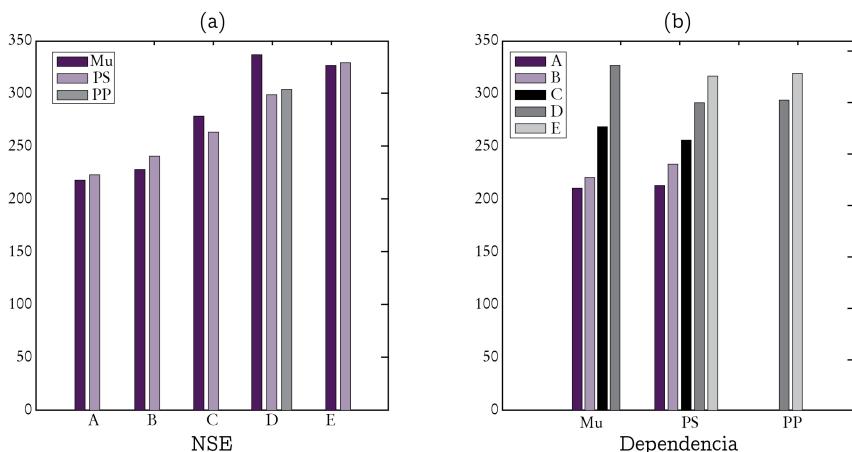


el promedio SIMCE: A igual nivel socioeconómico, los colegios particulares pagados tienen el mayor puntaje, seguido de los particulares subvencionados, salvo el caso de los colegios municipales de NSE D que obtienen el mejor puntaje promedio. Explique este caso.

TABLA 2.15. Promedio SIMCE según la dependencia y NSE

	A	B	C	D	E	Total
Municipal	217,8	227,8	277,0	335,2		235,7
Part. Subvencionado	219,9	239,9	263,4	298,2	325,1	256,0
Part. Pagado				302,7	327,6	324,7
Total	218,4	233,6	267,4	302,7	327,5	252,3

FIGURA 2.27. Promedio SIMCE según NSE y dependencia



Para terminar este capítulo, cuando se calcula un promedio, una mediana, una desviación estándar o se hace un gráfico, es importante tener presente el propósito para el cual se va a usar. En particular, la construcción de un gráfico debe ser:

- **fiable**: debe respetar lo que representan los datos;
- **leíble**: debe mostrar lo esencial de los datos, no agregar informaciones inútiles;
- **autosuficiente**: debe entenderse sin otras informaciones;
- **válido**: debe respetar las proporciones de las dimensiones, si quiere mostrar un círculo, debe verse un círculo y no una elipse, etc.

## 2.8 Resumen de la terminología

Población o universo: El conjunto de todos los objetos que se quiere estudiar.

Variable: Cantidad observada sobre los elementos de una población. Por ejemplo, la edad.

Variable cuantitativa: Variable que se puede medir numéricamente. Por ejemplos, la edad, el peso.

Variable continua: Variable cuantitativa que puede tomar cualquier valor de  $\mathbb{R}$  o de un intervalo de  $\mathbb{R}$ .

Variable discreta: Variable cuantitativa cuyos valores son aislados. Por ejemplo, el número de hijos.

Variable nominal: Variable que no se mide numéricamente, pero a través de "nombres". Por ejemplo, el color del pelo.

Variable ordinal: Variable nominal cuyos valores pueden ordenarse.

Categoría o modalidad: Los valores no numéricos tomados por una variable nominal o ordinal.

Frecuencia: Cantidad de observaciones tomadas por un valor de una variable.

Media: Es un valor que define el centro de una distribución.

Mediana: La cantidad de valores inferiores o iguales a la mediana es igual a la cantidad de valores superiores o iguales a la mediana.

Moda: El valor más frecuente que toma una variable.

Recorrido: La diferencia entre el valor mayor y el valor menor tomados por la variable.

Varianza: Es una medida de dispersión ligada a la media aritmética .

Desviación estándar: Es la raíz de la varianza.

Coeficiente de variación: Es el cociente de la desviación estándar y de la media aritmética.

## 2.9 Ejercicios

Los ejercicios con \* pueden utilizarse con los estudiantes de Enseñanza Media.

1. (\*) Consideremos estas poblaciones y variables:

Para cada caso, clasifique la o las variables según los cuatro tipos definidos en la sección 2.3.

2. (\*)<sup>6</sup>

Incremento de la criminalidad

El gráfico que figura a continuación se ha extraído del semanario de Zedlandia, El Noticario.

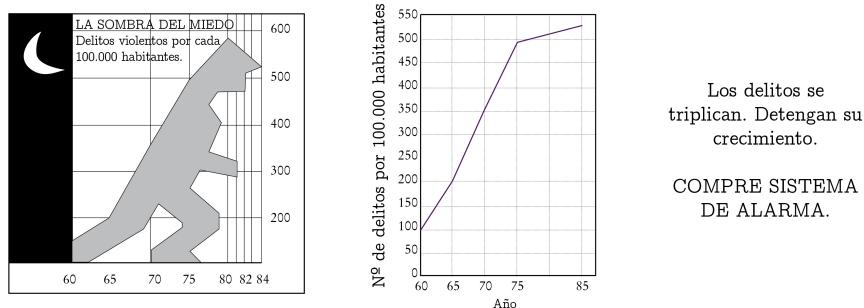
En él se muestra el número de delitos registrados por cada 100.000 habitantes, primero en intervalos de cinco años y luego en intervalos de un año.

- (a) ¿Cuántos delitos por cada 100.000 habitantes se registraron en 1960?

<sup>6</sup>Pregunta PISA 2006.

Población	Variables
1. Hogares chilenos	Ingreso del año 2007, comuna
2. Diputados del parlamento chileno en 2008	Partido político, votación con que fue electo
3. Comunas chilenas	Número de habitantes, superficie
4. Camiones	Carga máxima, número de ruedas
5. Hogares chilenos	Consumo anual de electricidad
6. Colegios en Chile	Promedio SIMCE 2do Medio
7. Galaxias	Número de estrellas
8. Niños chilenos menores de 5 años	Género, edad, peso, talla, color del pelo
9. Estudiantes de la Universidad de Chile	Carrera

- (b) Los fabricantes de sistemas de alarma recurrieron a estos mismos datos para elaborar el gráfico que aparece a continuación, ¿cómo y por qué elaboraron los fabricantes este gráfico?
- (c) A la policía no le hizo ninguna gracia el gráfico de los fabricantes de sistemas de alarma, porque quería demostrar que su lucha contra la delincuencia estaba resultando muy eficaz. Elabore un gráfico al que pueda recurrir la policía para demostrar que en los últimos tiempos se ha producido un descenso de la criminalidad.



3. Clasifique si los resultados siguientes provienen de la “estadística descriptiva” o de la “inferencia estadística”:

- (a) Pedro predice que el candidato Belair va a ganar la elección presidencial con un 53 % de los votos a partir de los resultados de 45 comunas.
- (b) El ecologista Sr. Lavados dice que cada pejerrey del lago Rapel contiene en promedio 400 unidades de mercurio.
- (c) En el Colegio de Llullay, el promedio de la PSU de Lenguaje fue de 550.
- (d) Se prevén 25 accidentes fatales para el próximo fin de semana largo.
- (e) El año pasado el 72 % de los trabajadores de la fábrica Acme perdieron al menos un día de trabajo por enfermedad laboral.
4. (\*) Sea la serie de datos: 3 5 5 6 7 8 8 9 9 10 10 11 12 12 12 15 18

- Si se reemplaza el 18 por 20, ¿cuál medida de posición central no cambia?
- Si se reemplaza el 18 por 20, ¿los cuartiles cambian?

5. (\*) A 20 alumnos de su curso se les pregunta el número de horas semanales que pasan frente al televisor. Las respuestas fueron: 11 12 12 9 5 11 7 9 10 11 10 8 8 9 10 9 14 13 12 7.

- Calcule el recorrido, la media, la mediana y la desviación estándar de los datos.
- ¿Qué ocurre con estas medidas si le suma tres a cada valor?
- ¿Qué ocurre si multiplica cada valor por dos?

6. (\*) Compare las distribuciones de frecuencias de las 6 notas de la Tabla 2.1 usando las herramientas vistas en este capítulo.

- (a) ¿Cuál de las 6 pruebas tiene el promedio más alto?
- (b) ¿Cuál de las pruebas tiene notas más concentradas?
- (c) ¿Cuál tiene una distribución más simétrica?
- (d) ¿Podemos decir que todos los alumnos del curso mejoraron su rendimiento en matemática entre la primera y la sexta prueba?
- (e) ¿Cuál es el alumno más destacado?

7. (\*) Un amigo norteamericano le regaló un termómetro de pared que muestra la temperatura en grados Fahrenheit (ver Figura 2.28). Usted observó la temperatura al mediodía durante 15 días y registró los siguientes resultados:

64 73 77 80 60 82 89 66 68 71 72 72 78 80 78

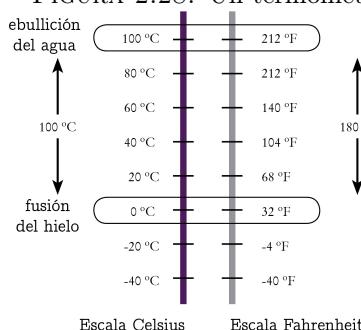
- (a) Sabiendo que la conversión de grados Centígrados a grados Fahrenheit es:

$$(^{\circ}F) = 1,8(^{\circ}C) + 32,$$

obtenga el promedio de las temperaturas en grados Centígrados.

- (b) Detectó que el termómetro no está bien calibrado y marca 5° Fahrenheit menos de lo real. ¿Cuál sería el promedio real en grados Centígrados?

FIGURA 2.28. Un termómetro



8. (\*) ¿Cuál es la mediana de las series de datos siguientes?

- (a) 5 ; 15 ; 17 ; 100 ; 105 ; 121 ; 2003.
- (b) 10 ; 3 ; -2 ; 5 ; 10 ; 12; 25 ; 4.
- (c) 12,5 ; 14 ; 5 ; 7 ; 9 ; 9 ; 4 ; 2,7 ; 8 ; 7.

9. (\*)

- Construye el diagrama de caja del conjunto de datos cuyo mínimo es 4, máximo es 30, el primer cuartil es 8, la median es 14 y el tercer cuartil es 22.
- Construye el diagrama de caja del conjunto de datos cuya distribución es simétrica, su mínimo es 2, su primer cuartil es 10 y su intervalo intercuartil es 10.

10. (\*) En la tabla de abajo, la primera columna es un identificador del individuo (niños, en nuestro caso) y las otras indican el nombre, la edad, el peso y la estatura.

Niño	Nombre	Edad (años)	Peso (kg)	Estatura (cm)
1	Julio	4	14,56	102,5
2	Alberto	5	16,35	112,0
3	Mariana	6	18,00	112,0
4	Carola	5	15,03	105,2
5	Rodrigo	6	17,20	109,5
6	Marcela	7	18,60	117,3
7	Juana	8	20,90	121,0
8	Raúl	6	15,30	106,3
9	Silvia	7	16,82	111,3
10	Anita	11	37,40	140,1
11	Patricio	10	26,44	131,1
12	Martín	9	29,55	130,2

- (a) Describa las columnas de la tabla.
- (b) Calcule los promedios, las desviaciones estándares y los coeficientes de variación de la edad, el peso y la estatura.
- (c) Si midiéramos la estatura de los alumnos en metros, ¿cómo cambiarían el promedio, la desviación estándar y el coeficiente de variación obtenidos en el punto anterior(b)?
- (d) Haga un gráfico de dispersión del peso y de la estatura. Agregue el género en el gráfico e interprete.

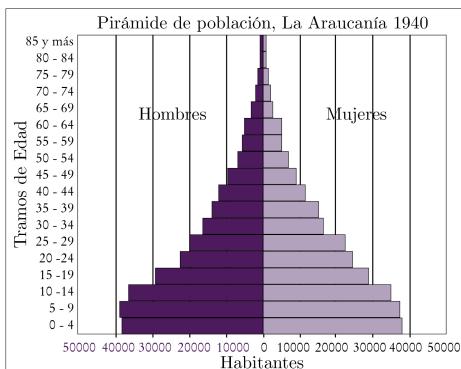
11. Según un estudio realizado por la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), la disponibilidad de agua por persona ha descendido bruscamente en un lapso de aproximadamente 50 años. La siguiente tabla señala la disponibilidad de agua en miles de metros cúbicos:

Región	1950	2000
África	17,8	4,8
Asia	7,6	2,9
Europa	5,9	4,5
Norte América	32,4	17,6
América Latina	72,1	22,8
Ex URSS	24,1	14,8
Oceanía	159,5	65,6

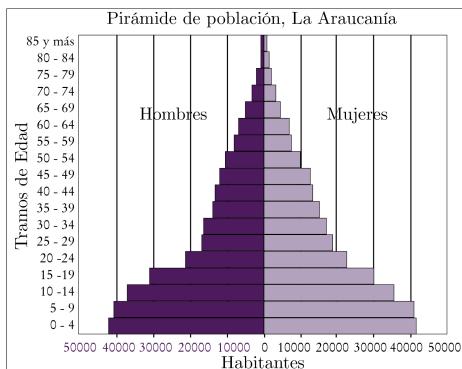
- (a) Haga un diagrama de barras que permita comparar la disponibilidad de agua en ambos años.
- (b) Calcule el porcentaje de descenso para cada región.
- (c) ¿Qué explicación le daría Ud. a estos descensos en la cantidad de agua per cápita?
12. Se muestran las pirámides de edades de la región de Araucanía para los años 1940, 1960, 1982 y 2002 (Figura 2.29)<sup>7</sup>.
- (a) ¿Qué muestran los cuatro gráficos?
- (b) ¿Qué piensa de la tendencia de la distribución etaria?
- (c) ¿Qué cree que pasará a futuro?
13. Para llegar a su trabajo, un automovilista recorre 10 km de un camino rural a una velocidad de 80 km/h. Cuando llega a la autopista recorre 50 km a la velocidad de 100 km/h y, finalmente, llega a la zona urbana donde disminuye la velocidad a 40 km/h durante 10 km.
- (a) ¿Cuál fue la velocidad promedio para llegar a su trabajo? Justifique su cálculo.
- (b) Compare con otras medias que podría calcular.
14. Una piscina puede vaciarse en 4 horas con una bomba a gas y en 6 horas con una bomba eléctrica.
- (a) ¿Cuánto tiempo tomará vaciar la piscina, si ponemos las dos bombas al mismo tiempo?
- (b) Calcule la media armónica de 6 y 4.
- (c) Muestre que la media armónica de dos números  $a$  y  $b$  puede escribirse como  $\frac{2ab}{a+b}$
- (d) Muestre que la media geométrica de dos números  $a$  y  $b$  es la media geométrica del promedio aritmético y de la media armónica.
15. Considere las dos series de datos siguientes:
- Serie(1): 10 2 6 5 9 8 5 0 8 4 6 8 9 7 2 4 9 .
- Serie(2): 10 2 6 5 9 8 5 0 8 4 6 8 9 7 2 4 9 9.
- (a) Ordene los datos en orden creciente de cada serie.

<sup>7</sup>Fuente: Observatorio Económico-Social de la Araucanía, Universidad de la Frontera

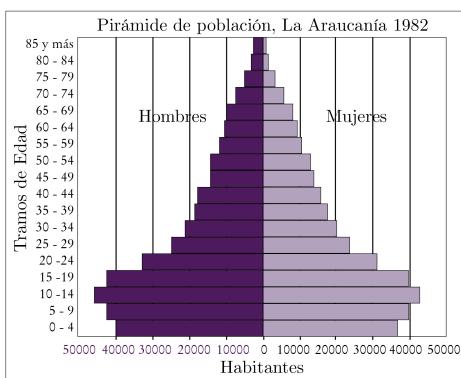
FIGURA 2.29. Pirámides de edades



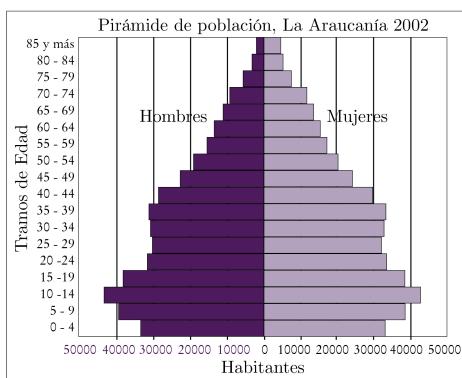
Fuente: Elaboración propia en base a Censo de Población, 1960.



Fuente: Elaboración propia en base a Censo de Población, 1960.



Fuente: Elaboración propia en base a Censo de Población, 1982.



Fuente: Elaboración propia en base a Censo de Población, 2002.

- (b) Calcule el promedio y la mediana de los datos de cada serie.
- (c) Calcule las expresiones 2.2 para los promedios y 2.3 para las medianas que encontraron.
- (d) Calcule las mismas expresiones tomando valores vecinos al promedio o a la mediana.
- (e) ¿Qué puede deducir respecto de la unicidad del mínimo de las expresiones 2.2 y 2.3 en el caso de un número de datos par o impar?
16. (\*) El restaurante Pizza Pino ofrece el servicio de entrega a domicilio, sin cargo, en un área con radio de 10 kilómetros. El gerente hizo tomar los tiempos de entrega durante un mes y obtuvo la siguiente información. Mínimo=13 min, Q1 (primer cuartil)=15 min, Q2 (mediana)=18 min; Q3 (tercer cuartil)=22 min, máximo=30 min.
- (a) Construya un diagrama de caja de los tiempos de entrega.

- (b) ¿Cuánto tiempo toma una entrega típica?  
 (c) ¿Dentro de qué amplitud de variación de los tiempos se efectúa la mayoría de las entregas?

17. A continuación se le presentan los gastos de electricidad (en millones de pesos) de dos sectores forestales conformados por varias empresas de la X región durante el mes de Junio de 2001.

Sector Norte	153	197	127	182	157	185	190
Sector Sur	167	145	149	206	132	128	168

- (a) Calcule el cuantil 55 % del Sector Norte y el tercer cuartil del Sector Sur. Interprételos.  
 (b) Determine el coeficiente de variación (desviación estándar/promedio) para cada sector. ¿Son parecidos? Concluya.  
 (c) En el Sector Norte las tarifas disminuyeron en un 3 % como resultado de una negociación con el proveedor. Determine ¿cuál es el nuevo promedio en las cuentas de electricidad?  
 (d) Realice un diagrama de cajas para comparar ambos sectores. Interprete.

18. Los 24 alumnos de un curso de 4to básico atribuyeron una nota a cada uno de sus 23 compañeros (Tabla 2.30). La nota (un número de 0 a 20) mide la afinidad que tiene para su compañero (0 si no tiene ninguna afinidad y 20 si tiene una afinidad total). Este tipo de tablas se llama “socio-matriz”. El curso es mixto y se indicó el género por medio de colores en los márgenes: morado para niña y blanco para niño. La manera de leer la tabla es muy importante. Las notas de una misma fila son las notas atribuidas por un alumno a sus compañeros. Las notas de una misma columna son las notas recibidas por un alumno de parte de sus compañeros. La alumna 1 atribuyó un 5 al compañero 2, un 12 al 3, etc. Esta niña recibió un 10 del compañero 2, un 15 del 3, etc. El total de la primera fila (299) es la suma de las notas atribuidas por la niña 1. El total de la columna 1 (349) es el total que la niña 1 recibió de parte de sus 23 compañeros.

Haga un diagrama en barras de todas las notas atribuidas. Haga el gráfico de dispersión de los 24 alumnos considerando en la abscisa *el total de notas recibidas* y en la ordenada *el total de notas atribuidas*. Trace la recta que va del punto (150,150) al punto (368,368).

- (a) ¿Cuáles son las notas atribuidas más frecuentes? Concluya.  
 (b) En el gráfico de dispersión, interprete qué significa que un alumno esté por debajo o por encima de la recta.  
 (c) ¿Hay diferentes afinidades entre niños y niñas?  
 (d) ¿Qué opinan del grupo los alumnos 22, 23, 1, 14 y 17?  
 (e) ¿Qué pasa con el niño 13?  
 (f) ¿Qué pasa con las niñas 12 y 16?

FIGURA 2.30. Socio-matriz

		Nota atribuida																								
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total
Nota recibida	1	5	12	14	15	3	18	10	10	10	14	14	5	20	17	15	19	17	10	16	12	16	15	12	299	
	2	10	17	16	14	20	17	19	13	17	11	13	8	10	6	1	17	17	16	17	13	17	17	19	325	
3	15	9	11	19	17	13	7	16	8	9	2	3	15	14	10	13	14	14	10	10	16	16	17	278		
4	15	12	11	10	17	12	16	13	15	12	10	8	8	11	11	13	13	17	16	14	14	15	18	301		
5	15	4	19	11	17	15	8	16	9	3	1	2	15	10	10	10	3	13	9	13	15	15	13	246		
6	9	19	18	13	18	10	5	12	2	1	5	2	12	3	17	15	15	11	11	12	16	14	242			
7	16	15	14	10	15	15	10	16	10	8	2	1	16	6	13	20	2	8	10	8	18	16	14	263		
8	17	18	15	13	16	18	12	15	16	17	13	11	16	10	14	16	15	12	19	18	18	17	18	354		
9	14	14	12	10	8	10	15	10	9	14	8	5	14	13	11	14	9	10	11	13	18	19	15	276		
10	19	10	15	12	10	15	13	20	15	12	8	12	17	12	3	12	12	10	20	14	15	15	15	306		
11	19	12	15	12	15	8	14	14	17	9	15	5	19	16	19	17	16	10	10	19	18	18	15	332		
12	20	15	14	17	16	13	13	10	17	13	14	13	19	10	12	11	19	8	14	14	18	18	9	327		
13	13	2	15	11	3	19	10	12	12	18	8	19	12	16	12	12	20	4	18	10	15	16	17	294		
14	20	5	13	14	15	5	18	16	10	10	14	13	5	17	16	17	13	10	16	11	15	15	13	301		
15	14	10	8	6	14	10	5	9	15	7	11	6	3	13	13	19	14	1	10	7	18	17	2	232		
16	10	10	5	3	6	6	8	7	9	6	15	12	1	1	11	10	8	10	4	5	11	13	8	3	181	
17	18	17	10	10	14	17	20	17	16	10	16	1	1	18	17	15	1	10	9	10	19	17	13	296		
18	15	12	10	10	3	16	10	13	10	10	11	20	17	15	18	10	10	10	13	19	14	14	10	290		
19	12	14	14	13	17	18	14	12	13	11	10	5	12	12	12	11	11	10	11	11	11	12	14	280		
20	18	9	11	16	6	15	13	20	17	19	12	8	13	18	19	10	11	11	14	14	19	17	15	325		
21	20	8	14	12	16	8	7	10	10	6	19	11	4	19	7	16	11	16	2	16	14	13	12	271		
22	14	13	13	12	13	10	16	12	18	12	10	12	12	14	18	9	15	10	9	13	10	19	18	302		
23	14	13	3	9	10	11	15	9	19	10	12	12	6	14	13	12	13	13	12	13	18	16	290			
24	12	15	16	19	15	15	11	12	13	13	4	12	13	12	12	7	12	11	14	13	11	17	17	296		
Total	349	261	304	274	288	303	299	278	322	250	258	218	165	329	296	253	318	281	234	299	286	368	362	312		

19. La primera tabla adjunta proporciona los números de alumnos por región y dependencia del colegio que rindieron la prueba SIMCE en 2006.

- (a) Utilizando una planilla Excel construya la tabla que permite ver si las regiones tienen las mismas proporciones de alumnos de colegios municipales, particulares pagados y particulares subvencionados. Interprete.
- (b) Interprete la segunda tabla que proporciona los promedios SIMCE por región y dependencia.

Región	Municipal	Part. pagado	Part. subvencionado	Total
1	3001	147	3902	7050
2	5114	610	1929	7653
3	2661	134	1715	4510
4	5616	247	4493	10356
5	9721	2118	12512	24351
6	7465	832	4951	13248
7	8433	531	6554	15518
8	17754	1353	11732	30839
9	6960	410	7840	15210
10	9779	721	6111	16611
11	595	0	870	1465
12	1213	171	803	2187
13	24995	10232	57395	92622
Total	103307	17506	120807	241620

Región	Municipal	Part. pagado	Part. subvencionado	Total
1	224	322	256	244
2	234	314	279	252
3	232	323	267	248
4	230	332	271	250
5	225	320	263	253
6	232	328	264	250
7	240	329	257	250
8	232	326	268	250
9	223	328	248	238
10	242	336	257	251
11	234		280	261
12	233	310	268	252
13	247	325	249	257
Total	236	325	256	252



## Capítulo 3: Introducción a la inferencia estadística



### 3.1 Motivación

En el capítulo anterior hablamos de distribuciones de frecuencias de variables medidas sobre un conjunto de observaciones, dando por entendido en forma implícita que esas observaciones constituyían la totalidad de la población.

A veces no podemos observar una variable sobre toda la población, pero podemos hacerlo en un subconjunto de ella, con la esperanza de que nuestras conclusiones sean generalizables a la totalidad. En estadística, el subconjunto se llama “**muestra**”.

A continuación presentamos algunos casos cuyo estudio se basa en una muestra.

- Para predecir los resultados de una elección presidencial, los encuestadores usan muestras de electores.
- La medición del nivel de pobreza en Chile necesita recoger muchos datos, usando un cuestionario complejo, que sería demasiado largo y caro aplicar en todos los hogares del país.
- La protección de las ballenas y otras especies marinas requiere conocer la cantidad de ejemplares existentes, pero es imposible contarlos a todos.
- Ciertas mediciones de control de calidad son destructivas (un ejemplo clásico es el de los fósforos, que hay que prender para comprobar que funcionan). En estos casos, obviamente sólo se puede ensayar una muestra.
- Para detectar anemia en un paciente, basta con algunas gotas de sangre (De hecho los médicos hablan de tomarle una “muestra” de sangre).
- Para saber si su sopa está bien aliñada, el cocinero prueba solamente una cucharada de sopa.

En la mayoría de los estudios científicos es imposible estudiar más de un fragmento del fenómeno considerado. Sería absurdo interrogar a todos los votantes antes de una elección o pescar todas las ballenas del océano. El fabricante no puede prender todos los fósforos que produce, el tecnólogo no puede extraer toda la sangre de un paciente, y el cocinero no se toma toda la sopa para probarla.

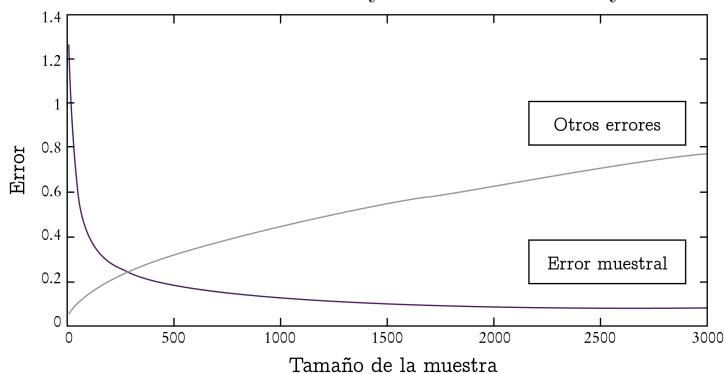
La pregunta natural es ¿cómo obtener una “**buena muestra**”? La respuesta, como veremos, no es simple y requiere la formalización previa de algunos conceptos. Sin embargo, podemos decir intuitivamente que una muestra es buena cuando constituye una “foto miniatura” de la población, con una resolución y nivel de detalle suficiente para nuestros fines. Obviamente, esto va a depender de lo que queremos saber sobre la población. No es lo mismo saber el nivel de desnutrición de los niños chilenos, para definir una política de salud, que saber cuántas horas diarias pasan frente al televisor, para definir una campaña publicitaria de gaseosas.

La **Inferencia Estadística** nos entrega un conjunto de métodos y técnicas para sacar conclusiones sobre una población a partir de los resultados obtenidos de una muestra, o para concebir una muestra que permita obtener esas conclusiones en forma satisfactoria. Al diseñar una muestra, tratamos de “controlar” el error muestral, que es el error que resulta de usar una parte de la población para concluir sobre la población entera, pero no hay que olvidar que los estudios complejos también son vulnerables a **errores no muestrales**, provenientes de las imperfecciones del proceso de medición.

El uso de una muestra tiene muchas ventajas, si está bien diseñada y la medición está bien administrada. Además de reducir los costos de levantamiento de datos, permite entregar información en menos tiempo y abarcar eventualmente más aspectos del problema. Muchas personas piensan que los resultados proporcionados por un censo son “exactos”, y que una muestra entrega resultados menos precisos, pero la verdad es que aunque los censos no tienen errores muestrales, son muy vulnerables a errores no muestrales, por la complejidad de su gestión. De hecho, una muestra bien administrada, con personal mejor seleccionado, capacitado y supervisado, puede entregar mediciones de mejor calidad que un censo (ver Figura 3.1).



FIGURA 3.1. Tamaño de la muestra y errores muestrales y no muestrales



### 3.2 Teoría de muestreo

La **Teoría de muestreo** es una colección de métodos particulares para extraer una muestra dependiendo de la situación.

#### 3.2.1 ¿Cuáles son las condiciones de un buen diseño muestral?

La muestra no se puede extraer de cualquier manera. Para hacer inferencias correctas a partir de ella es necesario saber cómo se obtuvo.

Cuando usamos los valores medidos en una muestra, para inferir una característica de la población de donde ésta proviene, decimos que este resultado es una **estimación** de dicha característica. Por ejemplo, si tenemos una muestra de hogares chilenos, el promedio del gasto anual en educación obtenido en la muestra podría ser una **estimación** del promedio del gasto en educación de todos los hogares chilenos.

La elección de la muestra debe asegurar una cierta objetividad. Para tener una idea de la distribución del nivel de pobreza de los hogares chilenos no sería adecuado restringir la muestra sólo a las comunas pobres del país. Si lo hicieramos, estaríamos sobreestimando (inflando), el nivel de pobreza. En otras palabras, nuestra estimación tendría un **sesgo**.

El punto de partida de un buen diseño muestral es el **marco muestral**, que es la lista de elementos de la población de interés. Los marcos muestrales pueden ser difíciles de obtener y, a veces, aunque se disponga de un marco completo, éste puede conducir a diseños poco prácticos. Para evaluar el nivel de pobreza a través de una encuesta de hogares, podríamos querer tener la lista completa y actualizada de todos los hogares chilenos, pero esa lista no existe. En el caso de la elección presidencial, aunque la lista de los electores inscritos está disponible (es el registro electoral), sería difícil ubicarlos para hacer una encuesta a partir de esta lista.

Otra condición fundamental de un diseño muestral objetivo es que **todo elemento de la población<sup>1</sup> debe tener la posibilidad de formar parte de la muestra.**

En resumen, el diseño y el análisis de muestras deben cumplir las siguientes condiciones:

1. Las estimaciones no deben tener *sesgo*: no deben sobreestimar o subestimar las características de la población que se pretende evaluar.
2. La muestra debe elegirse con objetividad: todo elemento de la población debe tener la posibilidad de ser elegido en la muestra. Además, ningún factor personal debería intervenir en la selección.
3. Para hacer inferencia hacia la población necesitamos una formalización matemática que permita estudiar las propiedades de la muestra, especialmente los errores inducidos por el muestreo.
4. Se deben controlar los errores muestrales.

Finalmente, elaborar un diseño muestral consiste en:

1. Construir el marco muestral para identificar la población.
2. Elegir el método de muestreo.
3. Elegir el tamaño de la muestra.

### 3.2.2 Muestreo aleatorio simple

Las condiciones citadas anteriormente para seleccionar una muestra se cumplen fácilmente dando un carácter aleatorio al muestreo y asignando a cada elemento de la población una probabilidad de selección **positiva**. Al analizar los datos obtenidos de una muestra aleatoria será necesario conocer esas probabilidades para obtener estimaciones inseguras.

*Se dice que un muestreo es aleatorio (o probabilístico, o científico), si todo elemento de la población tiene una probabilidad **no nula** y **conocida** de ser elegido en la muestra.*

Los elementos de la población pueden tener diferentes importancias. Para evaluar, por ejemplo, el crecimiento de la economía chilena mediante una muestra de empresas, es natural dar más importancia a las empresas grandes que a las chicas y otorgar probabilidades de selección dependientes del tamaño. No es el único caso que lleva a usar muestreo con probabilidades desiguales. Cuando las probabilidades de selección son desiguales, hay que conocerlas y tomarlas en cuenta en las fórmulas de estimaciones[12].

Como su nombre lo indica, el caso más simple de muestreo aleatorio es el llamado **muestreo aleatorio simple**:

---

<sup>1</sup>Los elementos de la población se llaman también **unidades estadísticas**.

*El muestreo aleatorio simple (m.a.s.) es un método de selección en el cual las unidades se eligen con la misma probabilidad. Se extraen las unidades una por una, de manera de que cada nueva unidad se obtiene del conjunto de las unidades no extraídas con equiprobabilidad.*

En la práctica habitual, los elementos de la muestra se eligen sin repetición. En otras palabras, la muestra es un subconjunto de la población. Por ejemplo, en una encuesta de hogares sería poco razonable entrevistar dos veces al mismo hogar<sup>2</sup>. Salvo que se especifique lo contrario, supondremos aquí que las muestras son sin repetición.

Para obtener una m.a.s., se extraen los elementos uno a uno, cada vez dándoles a todos los elementos remanentes la misma probabilidad de ser extraídos. Es como una lotería en que las fichas se van extrayendo sucesivamente de una bolsa que se revuelve cada vez.

Cuando el muestreo es aleatorio simple, todas las muestras posibles del mismo tamaño son equiprobables. Veamos el caso de una población finita con  $N$  elementos, hay  $\binom{N}{n}$  muestras posibles. Luego, cada muestra tiene una probabilidad igual a:

$$\frac{1}{\binom{N}{n}} = \frac{n}{N} \times \frac{n-1}{N-1} \times \frac{n-2}{N-2} \times \dots \times \frac{1}{N-n+1} = \frac{n!(N-n)!}{N!}.$$

En las salas de clase, el muestreo aleatorio simple puede ilustrarse usando un pequeño jarro transparente, en el cual se ponen entre 50 y 100 bolitas de dos colores diferentes, que representan, por ejemplo, dos candidatos a una elección (Figura 3.2). En el costado del jarro se hace una ventana que muestra algunas de las bolitas. Agitamos el jarro y consideramos a las bolitas visibles en la ventana como una muestra. Podemos, entonces, contar cuántas de cada color aparecen en la ventana, calcular la proporción de bolitas blancas en la muestra, y usar esa proporción como estimación de la proporción de bolitas blancas en todo el jarro. Se puede repetir el experimento varias veces, agitando el jarro para obtener distintas muestras, y calculando cada vez la proporción de bolitas blancas. Se puede, entonces, construir la distribución de frecuencias para mostrar la variabilidad de los resultados.

### 3.2.3 Otros tipos de muestreos

Como dijimos anteriormente, a menudo no se dispone de una lista de todos los elementos de la población (en otras palabras, no tenemos un marco muestral). Hasta puede suceder que ni siquiera se conozca el tamaño de la población, o que ésta sea

<sup>2</sup>Aunque los muestreos con repetición también son importantes desde el punto de vista académico, porque las fórmulas son algo más simples. En los experimentos de probabilidad, por ejemplo, se extrae una bolita de una urna y se le repone antes de extraer la siguiente.

FIGURA 3.2. Experimento con un jarro



infinita<sup>3</sup>. En estos casos no puede usarse el m.a.s. y hay que recurrir a otras técnicas. Más adelante volveremos a ocuparnos del m.a.s., pero antes daremos una breve descripción de otras técnicas de muestreo que se usan mucho en la práctica.

A menudo conviene clasificar las unidades estadísticas de la población en grupos y extraer una muestra independiente en cada grupo. Los grupos se llaman **estratos** y esta técnica se llama **muestreo estratificado**. Hay dos motivos distintos (y no necesariamente complementarios) para recurrir a la estratificación: el primero es disminuir el error de muestreo, por la vía de ganar control sobre la composición de la muestra; el segundo es obtener estimaciones de calidad comparable para cada uno de los estratos. La repartición de la muestra entre estratos dependerá del motivo: En las encuestas empresariales, que intentan medir producción y empleo, la estratificación se orienta principalmente hacia el primer objetivo, y para eso clasifican a las empresas según su tamaño, con el fin de elegir de preferencia a las más grandes. En una muestra de colegios de Enseñanza Media, en cambio, los usuarios podrían interesarse en comparar el rendimiento de los establecimientos privados y municipales, y los estratos van a definirse según la dependencia, de acuerdo con este objetivo.

En cualquier caso, la estratificación suele otorgar, de manera deliberada, probabilidades de selección diferentes a los distintos elementos de la población y eso habrá que tomarlo en cuenta al momento de analizar los datos. Los promedios o proporciones obtenidos directamente sobre la muestra ya no serán estimaciones insesgadas de

<sup>3</sup>Según su cantidad de elementos, las poblaciones se califican como *finitas* o *infinitas*. En la práctica las poblaciones son generalmente finitas, pero a veces muy numerosas. En el caso del experimento “Cara-Sello”, la población es infinita.

los valores poblacionales. Para obtener estimaciones sin sesgo, hay que usar ponderaciones.

Otra técnica usada en la mayoría de las encuestas de hogares es el muestreo **en dos etapas**: en lugar de elegir directamente una muestra de hogares, se elige primero una muestra de pequeñas unidades de área (generalmente, áreas de empadronamiento censal), y luego una muestra de hogares en cada una de las unidades de área elegidas. Esto resuelve el problema planteado por la inexistencia de una lista actualizada de todos los hogares del país, sólo se requiere una lista de unidades de área, que es mucho más fácil de obtener y mantener. Por otra parte, también permite abaratar los costos de transporte de los encuestadores entre un hogar y otro, pues los hogares de la muestra no estarán dispersos en todo el territorio, sino relativamente apiñados en las unidades de área elegidas en la primera etapa. El precio a pagar por estos beneficios es un mayor error muestral, que proviene precisamente de la conglomeración espacial de la muestra.

**El muestreo por conglomerados** es un caso particular del muestreo en dos o más etapas, en el cual, en la última etapa, se toman las unidades.

El **muestreo sistemático** consiste en seleccionar la muestra aplicando pasos uniformes sobre la lista ordenada de los elementos de la población. Por ejemplo, en la lista de los 2000 colegios de EM de Chile, se determina una muestra de 100 instituciones, seleccionando uno de cada 20, y eligiendo al azar el punto de partida. Si el orden de los elementos en la lista no tiene relación ninguna con los fenómenos que se van a estudiar (por ejemplo, si los colegios están en orden alfabético), este método es prácticamente equivalente al m.a.s., pero si el orden tiene alguna correlación con las variables de interés (por ejemplo, si los colegios se ordenan según los resultados de la SIMCE), es posible que los errores muestrales de una muestra sistemática sean algo menores que los de una m.a.s.

Para ilustrar el muestreo sistemático, supongamos que queremos elegir una muestra de 6 colegios, de entre los 30 colegios de la Tabla 3.1. Después de ordenar los colegios según algún criterio interesante (por ejemplo, la prueba SIMCE), se determina el paso de muestreo P (en este caso,  $P=30/6=5$ ) y se elige al azar un número entre 1 y P (por ejemplo, 2). Se toma, entonces, como muestra los colegios ubicados en la lista en las posiciones 2,  $2+5=7$ ,  $7+5=12$ ,  $12+5=17$ ,  $17+5=22$  y  $22+5=27$ .

Tanto el muestreo sistemático como el m.a.s otorgan la misma probabilidad de selección a cada elemento de la población. En otras palabras, las muestras sistemáticas y las muestras aleatorias simples son ambas equiprobables. Sin embargo, en el m.a.s las unidades se eligen independientemente unas de otras, cosa que no ocurre en el muestreo sistemático. En el ejemplo anterior, sólo elegimos aleatoriamente el primer elemento de la muestra, que resultó ser el colegio 2, y eso determinó directamente la elección de los colegios 7, 12, 17, etc.

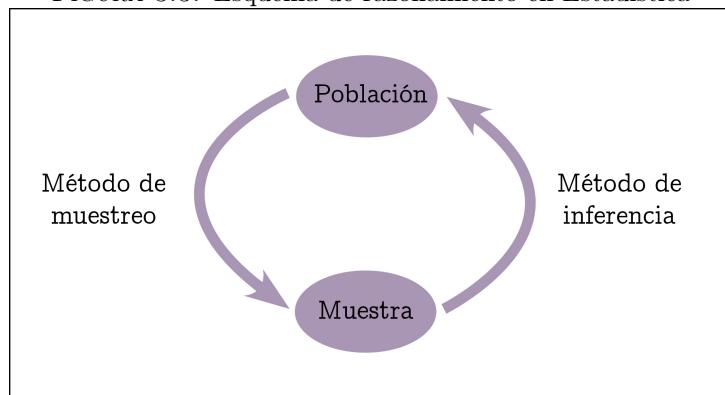
TABLA 3.1. Lista 30 colegios

ID	Colegio	SIMCE	Seleccionado en la muestra
1	Liceo Superior Gabriela Mistral	182	
2	Liceo Agrícola José Abelardo Núñez	190	✓
3	Liceo José Gutiérrez de la Fuente	193	
4	Liceo Politécnico Arica	194	
5	Liceo Elena Duvauchelle Cabezón	194	
6	Liceo de Huara	195	
7	Liceo Comercial y Técnico Arturo Prat	198	✓
8	Liceo Atenea	198	
9	Liceo Jovina Naranjo Fernández	205	
10	Liceo Luis Cruz Martínez	206	
11	Young School	215	
12	Liceo Domingo Latrille Loustaunou	215	✓
13	Liceo Particular Mixto Escasce	224	
14	Liceo Alcalde Sergio González Gutiérrez	224	
15	Liceo Granaderos	227	
16	Liceo Comercial Baldomero Wolnitzky	230	
17	Colegio Adventista de Iquique	242	✓
18	Colegio Alemán	248	
19	Liceo Domingo Santa María	250	
20	Colegio Int. Eduardo Frei Montalva	253	
21	Abraham Lincoln School	261	
22	Colegio Adventista	264	✓
23	Colegio San Marcos	273	
24	Colegio Italiano Santa Ana	293	
25	Liceo Octavio Palma Pérez	302	
26	North College	302	
27	Colegio San Jorge	313	✓
28	Liceo María Auxiliadora	318	
29	Junior College	323	
30	Liceo Academia Iquique	326	

### 3.3 Distribución en la población y distribución en la muestra

En el capítulo anterior vimos que para diseñar la muestra se necesitan algunas informaciones sobre la población (particularmente, un marco muestral). Los métodos de inferencia de los que hablaremos en este capítulo abordan el problema inverso: cómo obtener informaciones sobre la población a partir de la muestra (Figura 3.3).

FIGURA 3.3. Esquema de razonamiento en Estadística



Vamos a simular ahora casos donde conocemos la población, con el objeto de introducir ciertos conceptos y mostrar los elementos que intervienen en la precisión de las estimaciones.

Habitualmente estaremos interesados en estimar algunas de las características de la población (por ejemplo, la media, la varianza, la mediana, los cuartiles o quintiles, etc.), a partir de los valores observados en una muestra. Las características desconocidas de la población se llaman **parámetros**. Para estimar cada uno de esos parámetros, se usan funciones de los valores muestrales llamados **estimadores**.

Los valores muestrales encontrados en muestras aleatorias del mismo tamaño, son, en general, diferentes. En consecuencia, los estimadores tales como la media muestral, la varianza muestral, etc. van a variar dependiendo de la muestra con que se obtuvieron. En otras palabras, esos estimadores son variables aleatorias.

En esta monografía sólo nos vamos a preocupar por la estimación de la media, que es lejos el problema más importante en la práctica. Los métodos de inferencia que presentaremos supondrán conocida la varianza de la población (que en realidad también es desconocida y hay que estimarla), pero daremos también algunos resultados relativos a la varianza muestral en los ejemplos que vienen a continuación.

Veremos primero algunos ejemplos sencillos, de los cuales se deducen algunos resultados importantes sobre la media y la varianza muestrales de un muestreo aleatorio simple. Más específicamente, veremos cómo se relacionan las características de las muestras con las de la población.

### 3.3.1 Aprendamos de dos casos sencillos

Consideramos la pequeña isla de Santa Ana, compuesta de 4 aldeas  $\{A_1, A_2, A_3, A_4\}$  con 2, 3, 5 y 6 familias, respectivamente. Las aldeas de Santa Ana tienen en promedio 4,0 familias con una desviación estándar de 1,581 familias. San Blas, la isla vecina,

también tiene 4 aldeas  $\{B_1, B_2, B_3, B_4\}$ , y en ellas viven 1, 3, 5, y 7 familias, respectivamente. La aldeas de San Blas también tienen 4,0 familias en promedio, pero la desviación estándar de 2,238 familias es más grande que la de Santa Ana.

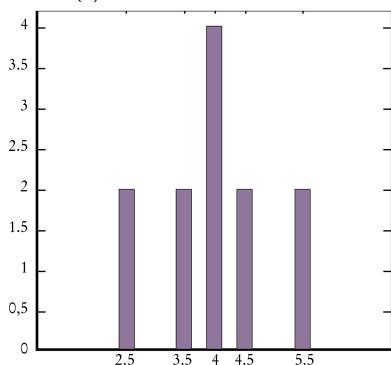
Saquemos, en ambas islas, todas las muestras posibles de dos aldeas, sin repetición. Hay 12 muestras posibles para cada isla. Para cada muestra de dos islas, calculemos la media, la varianza, la desviación estándar y el error muestral, que es la diferencia entre la media de la muestra y la media de la población (Tabla 3.2). En la Figura 3.4 se muestra la distribución de la media muestral de cada isla.

TABLA 3.2. Muestras de tamaño 2 sin repetición

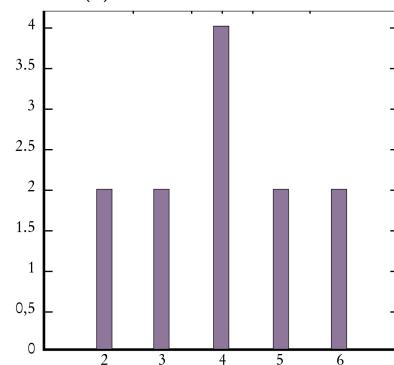
Santa Ana					San Blas				
Muestra	Media	Varianza	Desviación estándar	Error	Muestra	Media	Varianza	Desviación estándar	Error
$\{A_1, A_2\}$	2,5	0,25	0,5	-1,5	$\{B_1, B_2\}$	2,0	1,0	1,0	-2,0
$\{A_1, A_3\}$	3,5	2,25	1,5	-0,5	$\{B_1, B_3\}$	3,0	4,0	2,0	-1,0
$\{A_1, A_4\}$	4,0	4,0	2,0	0,0	$\{B_1, B_4\}$	4,0	9,0	3,0	0,0
$\{A_2, A_1\}$	2,5	0,25	0,5	-1,5	$\{B_2, B_1\}$	2,0	1,0	1,0	-2,0
$\{A_2, A_3\}$	4,0	1,0	1,0	0,0	$\{B_2, B_3\}$	4,0	1,0	1,0	0,0
$\{A_2, A_4\}$	4,5	2,25	1,5	0,5	$\{B_2, B_4\}$	5,0	4,0	2,0	1,0
$\{A_3, A_1\}$	3,5	2,25	1,5	-0,5	$\{B_3, B_1\}$	3,0	4,0	2,0	-1,0
$\{A_3, A_2\}$	4,0	1,0	1,0	0,0	$\{B_3, B_2\}$	4,0	1,0	1,0	0,0
$\{A_3, A_4\}$	5,5	0,25	0,5	1,5	$\{B_3, B_4\}$	6,0	1,0	1,0	2,0
$\{A_4, A_1\}$	4,0	4,0	2,0	0,0	$\{B_4, B_1\}$	4,0	9,0	3,0	0,0
$\{A_4, A_2\}$	4,5	2,25	1,5	0,5	$\{B_4, B_2\}$	5,0	4,0	2,0	1,0
$\{A_4, A_3\}$	5,5	0,25	0,5	1,5	$\{B_4, B_3\}$	6,0	1,0	1,0	2,0
Población	4,0	2,5	1,581		Población	4,0	5,0	2,236	

FIGURA 3.4. Distribución de frecuencias de la media muestral

(a) Media muestras Santa Ana



(b) Medida muestras San Blas



Las medias muestrales  $\{m_1, m_2, \dots, m_{12}\}$  son todos los valores posibles tomados por la variable aleatoria  $\bar{x}$ , la **media muestral** de todas las muestras de tamaño 2.

Lo que hemos hecho ilustra un hecho fundamental:

En un muestreo aleatorio, la media, la varianza, la mediana, el mínimo, el máximo, etc., de las diferentes muestras de mismo tamaño, son todas ellas variables aleatorias.

Pero, junto con señalar este hecho fundamental, es importante recordar que en la práctica del muestreo nunca vamos a elegir todas las muestras posibles; ni siquiera vamos a elegir unas pocas de ellas; en la práctica **siempre vamos a elegir una sola muestra**, y es a partir de esa única realización de las variables aleatorias que vamos a tener que inferir la media y otras características de la población.

Podemos calcular las características básicas de  $\bar{x}$  para cada isla. Comencemos con Santa Ana:

- La media (o **esperanza**) de  $\bar{x}$ :  $\mathbb{E}(\bar{x}) = \frac{1}{12} \sum_{i=1}^{12} m_i = 4,0$ .
- La varianza de  $\bar{x}$ :  $\text{Var}(\bar{x}) = \frac{1}{12} \sum_{i=1}^{12} (m_i - \mathbb{E}(\bar{x}))^2 = 0,8333$ .

Si llamamos  $X$  el número de familias en la aldea de Santa Ana,  $X$  puede tomar 4 valores (2, 3, 5 y 6) y la probabilidad asociada a cada valor en un muestreo equiprobable es  $1/4$ . Es la distribución de población de  $X$ :

$$\mathbb{P}(X = 2) = \frac{1}{4}; \quad \mathbb{P}(X = 3) = \frac{1}{4}; \quad \mathbb{P}(X = 5) = \frac{1}{4}; \quad \mathbb{P}(X = 6) = \frac{1}{4}.$$

La media muestral  $\bar{x}$  de las diferentes muestras de tamaño 2 tiene 5 valores posibles en Santa Ana: 2,5; 3,5; 4,0; 4,5; 5,5. La frecuencia de cada valor de la media es variable (Tabla 3.3). Para cada valor posible de la media muestral, la frecuencia relativa asociada indica la probabilidad de encontrar este valor de la media en una muestra aleatoria sin reemplazo de tamaño 2. Aquí, tenemos mayor probabilidad de encontrar una media muestral  $\bar{x}$  igual a 4, que los otros valores posibles que tienen una probabilidad de  $1/6$ . Observe que el valor más probable 4 es el valor real de la media en la población. Podemos escribir con el lenguaje de las probabilidades:

$$\mathbb{P}(\bar{x} = 2,5) = \frac{1}{6}; \quad \mathbb{P}(\bar{x} = 3,5) = \frac{1}{6}; \quad \mathbb{P}(\bar{x} = 4) = \frac{2}{6}; \quad \mathbb{P}(\bar{x} = 4,5) = \frac{1}{6}; \quad \mathbb{P}(\bar{x} = 5,5) = \frac{1}{6}.$$

TABLA 3.3. Frecuencias de los valores de la media muestral

	Santa Ana						San Blas					
	2,5	3,5	4,0	4,5	5,5	Total	2	3	4	5	6	Total
Media	2	2	4	2	2	12	2	2	4	2	2	12
Frecuencia absoluta	2	2	4	2	2	12	2	2	4	2	2	12
Frecuencia relativa	$1/6$	$1/6$	$1/3$	$1/6$	$1/6$	1	$1/6$	$1/6$	$1/3$	$1/6$	$1/6$	1

Se pueden calcular la esperanza y varianza de la v.a.  $\bar{x}$  a partir de la distribución de probabilidad  $\bar{x}$ . Para la isla de Santa Ana:

- La esperanza de  $\bar{x}$  es la suma ponderada de todos los valores que puede tomar  $\bar{x}$ , donde la ponderación de un valor es la probabilidad asociada a este valor:

$$\mathbb{E}(\bar{x}) = \frac{1}{6} \times 2,5 + \frac{1}{6} \times 3,5 + \frac{2}{6} \times 4 + \frac{1}{6} \times 4,5 + \frac{1}{6} \times 5,5 = 4,0.$$

- La varianza de  $\bar{x}$  es la suma ponderada de los cuadrados de todas las diferencias entre cada valor que puede tomar  $\bar{x}$  y su esperanza  $\mathbb{E}(\bar{x})$ :

$$\text{Var}(\bar{x}) = \frac{1}{6} \times (2,5 - 4,0)^2 + \frac{1}{6} \times (3,5 - 4,0)^2 + \frac{2}{6} \times (4,0 - 4,0)^2 + \frac{1}{6} \times (4,5 - 4,0)^2 + \frac{1}{6} \times (5,5 - 4,0)^2 = 0,8333$$

Se deja como ejercicio efectuar los mismos cálculos para la isla de San Blas.

La esperanza y varianza de  $\bar{x}$  nos dicen algo sobre la distribución de  $\bar{x}$ , pero también vale la pena observar otras características de la distribución muestral de  $\bar{x}$ . Por ejemplo, en el caso de Santa Ana, ¿cuál es la probabilidad de que  $\bar{x}$  sea más pequeña que 4? De la tabla de frecuencias 3.3 se deduce que esa probabilidad es 1/3.

En este caso, obtuvimos todos estos resultados por la simple vía de construir todas las muestras posibles. Sin embargo, como dijimos antes, en la práctica sólo podemos extraer una única muestra. Felizmente, existen fórmulas para calcular la esperanza y la varianza de  $\bar{x}$ , dependiendo del tipo de muestreo utilizado. Existen, además, modelos teóricos para la distribución de  $\bar{x}$ , que permiten aproximar muy bien la probabilidad de que  $\bar{x}$  no supere un valor dado cualquiera. El modelo teórico más utilizado es el modelo **Normal** (Ver 3.4.1). Veremos más adelante que el modelo Normal tiene su justificación en el **Teorema Central del Límite** (Ver 3.4.2).

### 3.3.2 Estimación de la media de la población: sesgo y precisión

En un muestreo aleatorio simple, si  $x_1, x_2, \dots, x_n$  son los valores de una muestra de tamaño  $n$  obtenida a partir de una población de tamaño  $N$ , con media  $\mu$  y varianza  $\sigma^2$ , entonces, la teoría de probabilidades permite deducir que la media y la varianza

de  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  son:

$$\mathbb{E}(\bar{x}) = \mu \quad \text{y} \quad \text{Var}(\bar{x}) = \frac{N - n}{N - 1} \frac{\sigma^2}{n}.$$

La formula nos dice que, en **promedio**, la media muestral es igual a la media de la población. En otras palabras, en una m.a.s. la media muestral no tiene **sesgo**: la diferencia entre la media del estimador ( $\mathbb{E}(\bar{x})$ ) y el parámetro ( $\mu$ ) es nula. Generalmente, no conocemos la media poblacional  $\mu$  (si la conocieramos, no tendría mucho sentido recurrir al muestreo para estimarla), pero la varianza de  $\bar{x}$  nos indica la magnitud de la variabilidad de los valores posibles de la media muestral; o sea, qué tan **preciso** es  $\bar{x}$  como estimador de  $\mu$ . Si la varianza es pequeña, podemos estar razonablemente confiados de que el valor de  $\bar{x}$  en nuestra muestra va a estar cercano a  $\mu$ . Es por eso que la estimación de  $\mu$ , por medio de  $\bar{x}$ , tiene que venir acompañada de la desviación

estándar de  $\bar{x}$ . Esa desviación se llama **error estándar** y está dada por:

$$e(\bar{x}) = \sqrt{\text{Var}(\bar{x})} = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{\sigma^2}{n}}. \quad (3.1)$$

Observemos que  $e(\bar{x})$  depende de tres elementos: el tamaño  $N$  de la población, el tamaño  $n$  de la muestra y la varianza  $\sigma^2$  de la población.

La formula 3.1 muestra que, en dos poblaciones de tamaños similares pero con distintas varianzas, la población con mayor varianza necesitará una muestra más grande que la de menor varianza, para que las dos tengan el mismo error estándar. En la Figura 3.5(a) se muestra cómo varía el error estándar de una media en función de  $n$ , para poblaciones de tamaño grande con varianzas iguales a 2, 4 y 20. Analice el gráfico.

Respecto del tamaño de la población, un error muy común es creer que las poblaciones grandes requieren muestras mayores que las poblaciones pequeñas. La mayoría cree que si dos poblaciones tienen la misma varianza y si una es dos veces más grande que la otra, entonces, la más grande requerirá de una muestra dos veces más grande que la más pequeña, para que las dos muestras tengan el mismo error estándar. ¡Esto es falso! Basta mirar la formula 3.1, para constatar que el tamaño de la población  $N$  sólo aparece en el término  $\sqrt{\frac{N-n}{N-1}}$ , que es muy cercano a 1 cuando el tamaño  $N$  de la población es grande con respecto del tamaño  $n$  de la muestra; que es lo que generalmente sucede en la práctica. En otras palabras, el tamaño de la población influye muy poco sobre la precisión.

En la figura 3.5(b) se puede ver cómo varía el tamaño de la muestra necesaria para obtener un error estándar de 0,05 en función del tamaño de la población, para poblaciones con varianzas 2 y 5. En ambos casos, el tamaño de la muestra necesaria se estabiliza cuando el tamaño de la población supera algunos miles.

Cuando el tamaño de la población es suficientemente grande, la fórmula 3.1 se reduce a:

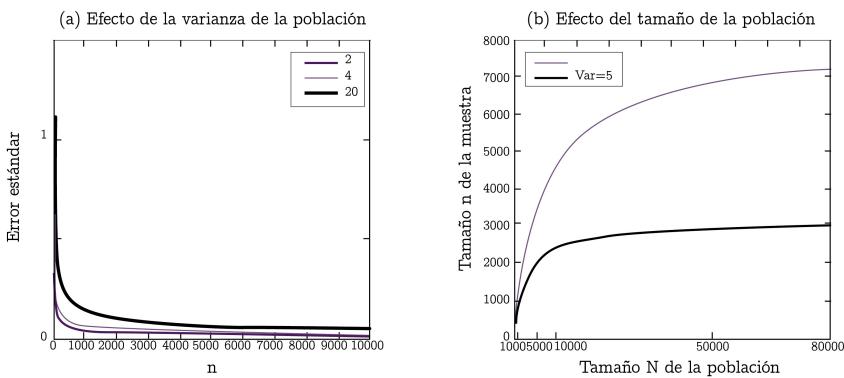
$$e(\bar{x}) = \sqrt{\text{Var}(\bar{x})} = \sqrt{\frac{\sigma^2}{n}}. \quad (3.2)$$

Notas:

1. En el caso de un muestreo aleatorio simple con reemplazo, el error estándar también está dado por la ecuación 3.2.
2. En poblaciones grandes, es equivalente hablar de muestreo con o sin reemplazo.
3. En poblaciones pequeñas, el muestreo sin reemplazo tiene un error estándar menor que el muestreo con reemplazo.

En resumen, el tamaño de la muestra es el factor más importante en la calidad de la inferencia hacia la población. La varianza de la población y el error estándar que estamos dispuestos a aceptar son los elementos más importante para elegir el

FIGURA 3.5. Error estándar de la media y tamaño de la muestra



tamaño de la muestra. El tamaño de la población sólo tiene importancia cuando ésta es pequeña.

Antes de abandonar este tema, se impone una nota de advertencia. Cuando decimos que la media muestral proporciona un estimador insesgado de la media poblacional, nos referimos sólo a los efectos del error muestral, que es el error que resulta de no observar la población entera sino una muestra; pero no debemos olvidar que las estimaciones también pueden verse afectadas por errores no muestrales. Si quisieramos estimar la velocidad promedio de los vehículos que viajan de Santiago a Valparaíso, promediando las velocidades medidas para una muestra de vehículos con un radar como el que usan los Carabineros, sólo podríamos decir que nuestra estimación no tiene sesgo muestral, pero si el aparato está mal calibrado, vamos a tener una sobre o subestimación, por buena que sea la muestra que elijamos.

La noción de **precisión** resume los efectos del error muestral y del sesgo por medio de un indicador combinado llamado **error cuadrático medio**. Si  $\hat{\theta}$  es un estimador de un parámetro  $\theta$  con sesgo  $\delta = \mathbb{E}(\hat{\theta}) - \theta$ , el error cuadrático medio está dado por:

$$\mathbb{E}\{(\hat{\theta} - \theta)^2\} = \mathbb{E}\{(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\} + (\mathbb{E}(\hat{\theta}) - \theta)^2 = \text{Var}(\hat{\theta}) + \delta^2.$$

### 3.4 Distribuciones teóricas

A continuación presentamos dos distribuciones de probabilidad importantes en Estadística: La distribución Normal para variables reales y la distribución de Bernoulli para variables binarias, así como el Teorema Central del Límite.

#### 3.4.1 La distribución Normal

La distribución Normal fue introducida por primera vez en 1733 por Abraham de Moivre en un artículo. Sus resultados fueron extendidos al mismo tiempo por Carl Friedrich Gauss (1777-1855), Adrien-Marie Legendre (1752-1833), y Pierre-Simón

Laplace (1749-1827), dentro de un estudio importante: el análisis numérico de los errores de mediciones en física y astronomía. Buscaban resolver el problema de cómo determinar el mejor valor leído por un instrumento que entrega diferentes mediciones del mismo fenómeno.

En efecto, si tenemos  $n$  mediciones de un mismo fenómeno  $\{x_1, x_2, \dots, x_n\}$ , deberíamos tener  $x_1 = x_2 = \dots = x_n$  si no tuviéramos errores. En su anexo de “Nuevos métodos para la determinación de las órbitas de los cometas”, Legendre propone, en 1805, el importante método de los Mínimos Cuadrados que permite determinar un valor único  $z$  de la medición de manera que una función de los errores sea mínima:

$$\min_z \sum_{i=1}^n (x_i - z)^2. \quad (3.3)$$

La solución es el promedio de las mediciones (derive la expresión 3.3 con respecto de  $z$  para comprobarlo).

Gauss, que afirmó haber utilizado este método desde 1794, lo justificó rigurosamente en 1809 asumiendo una distribución Normal de los errores, aunque la distribución de los errores fue estudiada mucho antes por Thomas Simpson (1710-1761). En este estudio, Simpson no usó específicamente la distribución Normal, pero hizo los supuestos que la distribución de los errores tenía que ser simétrica y que la probabilidad de errores pequeños debería ser mayor que la de los errores grandes.

En 1840, Sir Francis Galton (1822-1911), primo de Charles Darwin, partió de una distribución discreta y la fue refinando hasta llegar, en 1857, a una distribución continua muy parecida a la distribución normal (ver la distribución Binomial en el siguiente párrafo).

El nombre de “Campana de Gauss” viene de Esprit Jouffret (1872) y el nombre de “distribución Normal” viene de Galton (1875). (Figura 3.6)

*La distribución normal es la ley en la cual todo el mundo cree:*

*Los experimentadores creen, que es un teorema de la Matemática,  
los matemáticos, que es un hecho experimental.*

*El astrónomo Lippman.*

La función  $\phi(x)$  correspondiente a la campana de Gauss llamada **función de densidad** de la distribución normal de media 0 y varianza 1 ( $\mathcal{N}(0, 1)$ ), (también llamada distribución Gaussiana), tiene la expresión:

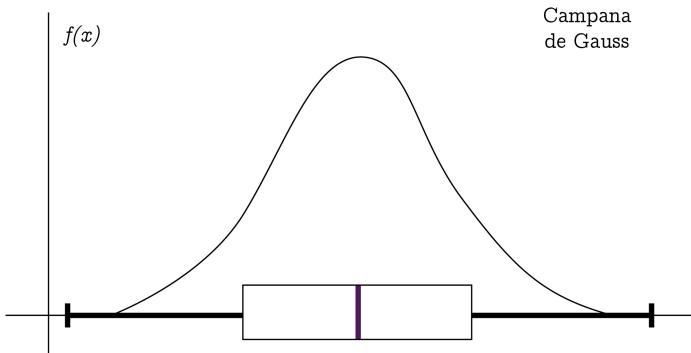
$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}x^2}, \quad \forall x \in R$$

Esta función es simétrica con respecto de 0, que es la moda de la distribución. La densidad disminuye paulatinamente en los extremos. La simetría se utiliza, a menudo, en los cálculos de probabilidad.

Si  $X \sim \mathcal{N}(0, 1)$ , a partir de la función se puede calcular probabilidades:

- $\mathbb{P}(X \leq 0) = \int_{-\infty}^0 \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}x^2} dx = \frac{1}{2}.$

FIGURA 3.6. La campana de Gauss



- $\mathbb{P}(X \geq 0) = \int_0^{\infty} \frac{1}{\sqrt{2}} e^{-\frac{1}{2}x^2} dx = \frac{1}{2}$ .
- $\mathbb{P}(X \leq 0,5) = \int_{-\infty}^{0,5} \frac{1}{\sqrt{2}} e^{-\frac{1}{2}x^2} dx = 0,6915$ .
- $\mathbb{P}(X \geq 1,96) = \int_{1,96}^{\infty} \frac{1}{\sqrt{2}} e^{-\frac{1}{2}x^2} dx = 0,025$ .
- Más generalmente,  $\mathbb{P}(X \leq x) = \int_{-\infty}^x \phi(x)dx = \mathbb{P}(X \geq -x)$

Estos cálculos serán la base de la construcción de un intervalo de confianza y de la teoría de tests de hipótesis.

La distribución Normal representa una familia de distribuciones que se obtienen a partir de la Normal Estandarizada ( $\mathcal{N}(0, 1)$ ). Se dice que  $X$  es una distribución Normal de media  $\mu$  y desviación estándar  $\sigma$  ( $X \sim \mathcal{N}(\mu, \sigma)$ ), si y solo si  $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$ . O sea, toda transformación lineal de una variable  $\mathcal{N}(0, 1)$  es una variable Normal también, pero su media y su varianza cambian. La expresión de la función de densidad de una distribución Normal de media  $\mu$  y desviación estándar  $\sigma$  ( $X \sim \mathcal{N}(\mu, \sigma)$ ) es simétrica con respecto de  $\mu$ :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad \forall x \in R, \quad \mu \in R, \quad \sigma > 0.$$

La función  $f(x)$  es simétrica con respecto de  $\mu$ .

Se tienen algunas propiedades importantes en relación con la distribución Normal.

- Si  $X \sim \mathcal{N}(\mu, \sigma)$  y  $b$  es un escalar constante, entonces  $X + b \sim \mathcal{N}(\mu + b, \sigma)$ .

La media se traslada de  $b$ , pero la varianza queda invariante.

- (b) Si  $X \sim \mathcal{N}(\mu, \sigma)$  y  $a$  es un escalar constante, entonces  $aX \sim \mathcal{N}(a\mu, |a|\sigma)$ . Observe la transformación de la media y de la varianza.
- (c) Si  $X \sim \mathcal{N}(\mu, \sigma)$  y  $a$  y  $b$  son escalares constantes, entonces  $aX + b \sim \mathcal{N}(a\mu + b, |a|\sigma)$ .
- (d) Sean  $X \sim \mathcal{N}(\mu, \sigma)$  y  $Y \sim \mathcal{N}(\nu, \tau)$ . Si  $X$  e  $Y$  son independientes (la distribución de una no depende de la otra), entonces  $X + Y \sim \mathcal{N}(\mu + \nu, \sqrt{\sigma^2 + \tau^2})$ .

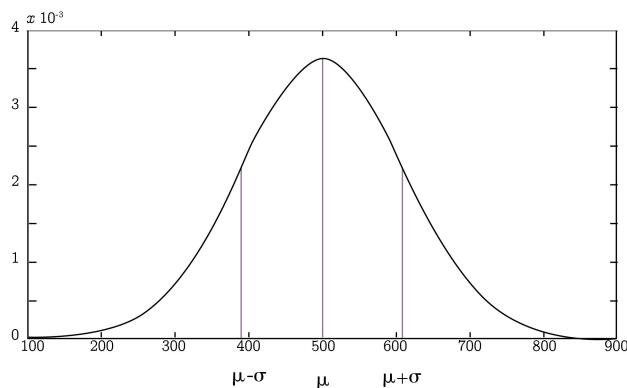
El histograma Figura 2.8 de la distribución de la PSU en el capítulo 2 tiene la forma de la campana de Gauss: hay cierta simetría, hay pocos puntajes en los extremos y un aumento paulatino hasta llegar a la parte central del recorrido, donde está la mayoría de ellos. Se puede aproximar la distribución empírica a la distribución teórica  $\mathcal{N}(500, 110)$ , donde la media de la PSU es 500 y 110 la desviación estándar.

Un resultado fundamental relativo a la media muestral de una distribución de población:

**Proposición 3.1.** *Si  $X_1, X_2, \dots, X_n$  es una m.a.s. de una variable  $X \sim \mathcal{N}(\mu, \sigma)$ , la media muestral  $\frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \sqrt{(\frac{N-n}{N-1}) \frac{\sigma^2}{n}})$  si la población es de tamaño  $N$  y  $\frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$  si la población es considerada como infinita o si se usa un muestreo con reemplazo.*

Hay dos puntos de inflexión en la curva de la función de densidad  $\mathcal{N}(\mu, \sigma)$ , que son situados a ambos lados de la media  $\mu$  a la distancia  $\sigma$  de ella. Son  $\mu - \sigma$  y  $\mu + \sigma$  (Figura 3.7).

FIGURA 3.7. Distribución  $\mathcal{N}(\mu, \sigma)$



### 3.4.2 Teorema Central del Límite

Un teorema fundamental de Probabilidad es el Teorema Central del Límite. Éste indica que, bajo condiciones muy generales, la distribución de la suma de variables aleatorias tiende a una distribución Normal, cuando la cantidad de variables es muy grande.

La versión más simple del teorema se basa en el concepto de variables aleatorias independientes. Se dice que dos v.a son independientes, si los valores que toma una variable no afectan las probabilidades asociadas a los valores que toma la otra y recíprocamente.

**Teorema 3.2.** *Sea  $\{X_1, X_2, \dots, X_n, \dots\}$  un conjunto de variables aleatorias “independientes” de misma distribución, de esperanza  $\mu$  y de varianza  $\sigma^2$  finitas. Sean*

$$\forall n \in \mathbb{N}^*, \quad \bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

y

$$Z_n = \frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu).$$

Cuando  $n$  crece, la distribución de  $Z_n$  converge, entonces, hacia la distribución normal de media nula y varianza igual a 1.

El Teorema Central del Límite ofrece, de este modo, una buena aproximación para la media muestral de una m.a.s. de una población de media  $\mu$  y varianza  $\sigma^2$ , una distribución normal  $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ , cuando el tamaño de la muestra es suficientemente grande.

Anteriormente vimos que es importante poder calcular probabilidades de la distribución de la media muestral u otra característica de la muestra. Si bien, a partir de una sola muestra es imposible, se puede usar la distribución normal como aproximación razonable en muchas situaciones cuando la muestra no es demasiado pequeña.

La importancia en estadística de la distribución normal se debe al Teorema Central del Límite. En muchas situaciones se podrá modelar una distribución empírica o muestral con tal distribución.

### 3.4.3 La distribución de Bernoulli

La distribución de Bernoulli corresponde al caso más simple de distribución discreta:  $X$  toma dos valores 1 o 0 con  $\mathbb{P}(X = 1) = p$  y  $\mathbb{P}(X = 0) = 1 - p$ , donde  $p$  es el parámetro de la distribución de Bernoulli ( $X \sim \mathcal{B}(p)$ ).

Mediante una distribución de Bernoulli podemos modelar muchas situaciones. Si, en el lanzamiento de una moneda equilibrada, definamos  $X = 1$  cuando sale “cara” y  $X = 0$  cuando sale “sello”,  $X$  tiene una distribución de Bernoulli de parámetro  $p = \frac{1}{2}$ .

En una elección presidencial a la cual se presentan dos candidatos, “Rojo” y “Valverde”, si definimos  $X = 1$  cuando un votante elegido aleatoriamente vota por “Valverde” y  $X = 0$  cuando vota por “Rojo”.  $X$  tiene una distribución de Bernoulli de

parámetro  $p$  igual a la probabilidad que un votante elegido aleatoriamente vota por “Valverde”.

La esperanza y la varianza de  $X$  son:

$$E(X) = 1 \times p + 0 \times (1-p) = p, \quad \text{Var}(X) = (1 - E(X))^2 \times p + (0 - E(X))^2 \times (1-p) = p(1-p)$$

Sea ahora los resultados de una encuesta de la intención de votos de la elección presidencial con un muestreo m.a.s. de tamaño  $n$  y  $\{x_1, x_2, \dots, x_n\}$  los valores muestrales ( $x_i = 1$ , si el voto es por “Valverde” y  $x_i = 0$  si es para “Rojo”). La proporción de votos  $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$  de la muestra es una estimación de la proporción  $p$  de votos de la elección general<sup>4</sup>. Es igual a la media de las variables binarias  $x_i$ .

En el caso de m.a.s. con reemplazo,  $\sum_{i=1}^n x_i$  es una suma de v.a. Bernoulli de mismo parámetro e independientes entre sí. La suma sigue una distribución llamada **distribución Binomial** ( $Binomial(n, p)$ ), cuya esperanza es igual a  $np$  y varianza a  $np(1-p)$ . Si  $Y$  sigue una distribución  $Binomial(n, p)$ :

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

En el caso de una m.a.s. sin remplazo con población finita  $N$ , tiene una distribución que ya no es una binomial<sup>5</sup>, pero su esperanza es  $np$  y su varianza es  $\frac{N-n}{N-1} np(1-p)$ .

En m.a.s. con o sin reemplazo, para  $n$  grande,  $\sum_{i=1}^n x_i$  tiene una distribución que se approxima a una distribución normal de media  $np$  y varianza  $np(1-p)$   
 $(\sum_{i=1}^n x_i \sim \mathcal{N}(np, \sqrt{np(1-p)}))$ . La proporción muestral  $\frac{1}{n} \sum_{i=1}^n x_i$  se approxima también a una distribución normal de media  $p$  y varianza  $\frac{p(1-p)}{n}$  ( $\frac{1}{n} \sum_{i=1}^n x_i \sim \mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}})$ ).

Galton inventó una máquina llamada **quincunx** o máquina de Galton, que permite ilustrar este último resultado. En esta máquina, se deja caer bolitas que van buscando su camino entre clavos regularmente repartidos en un triángulo. Cada vez que una bolita golpea un clavo va a la izquierda o derecha con igual probabilidad ( $1/2$ ). Al final las bolitas se acumulan en ranuras puestas debajo del triángulo. Cada golpe en un clavo corresponde a una distribución de Bernoulli ( $\mathcal{B}(1/2)$ ) y la cantidad de bolitas en las ranuras corresponde a una distribución Binomial. El experimento de Galton ilustra entonces el teorema Central del Límite. La aproximación a una distribución Normal será mejor mientras hay más filas de clavos.

---

<sup>4</sup>Se usa frecuentemente el sombrero sobre el parámetro para designar a un estimador del parámetro.

<sup>5</sup>Es una distribución hipergeométrica

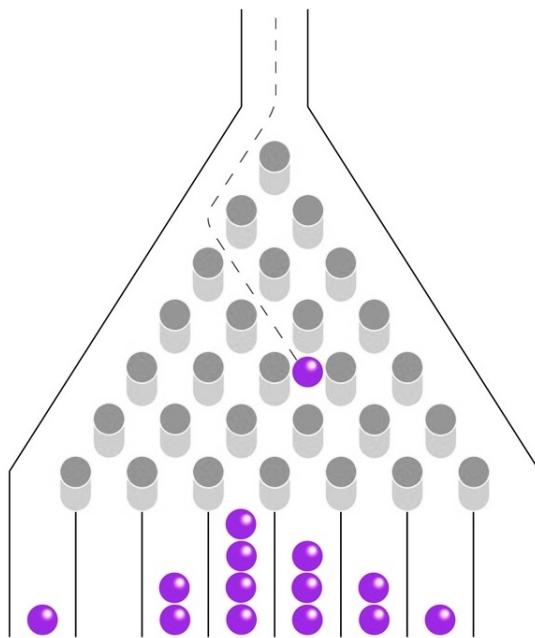
¿Cómo saber cuándo se asemeja la distribución binomial a la distribución Normal?  
La distribución binomial se asemeja a la distribución normal siempre que  $np(1 - p)$  es grande (mayor que 10).

Con sus estudiantes podrían intentar fabricar una máquina de Galton, si esto no es posible, en Internet se encuentran simuladores

“<http://www.math.uah.edu/stat/applets/GaltonBoardExperiment.xhtml>”.

### *Quincunx*

*En 1840, Sir Francis Galton (1822-1911), primo de Charles Darwin, partió de una distribución discreta y la fue refinando hasta llegar en 1857 a una distribución continua muy parecida a la distribución normal. Galton inventó incluso una máquina llamada quincunx o máquina de Galton, que permite ilustrar la distribución normal.*



### **3.5 Intervalos de confianza**

En la sección anterior, estimamos la media  $\mu$  de una población, usando como estimador la media muestral  $\bar{x}$  de un m.a.s, y dimos una fórmula para calcular la varianza de la media muestral (recordando que es una v.a.). Esta varianza es útil para comparar estimadores, pero todavía no sabemos cómo usarla para medir la precisión de la

estimación, o sea, para decir qué tan confiable es una media muestral como estimador de la media de la población.

La precisión de una estimación se hace construyendo intervalos con probabilidades conocidas de contener el verdadero valor de la media de la población. Este método se conoce con el nombre de **estimación por intervalos**. Para cada probabilidad que elijamos vamos a tener un intervalo de confianza distinto. Mientras mayor sea esa probabilidad, más amplio va a ser el intervalo, y aunque casi nunca vamos a poder definir un intervalo que nos dé la garantía absoluta de contener el valor verdadero, podremos en general definir intervalos que nos den una confianza razonablemente alta de hacerlo. Veremos en esta sección cómo definir intervalos de confianza para la media de una población, suponiendo conocida su varianza.

### 3.5.1 Media

Para introducir la estimación por intervalos, vamos a suponer que conocemos la distribución de la medición  $X$  en la población, por ejemplo  $X \sim \mathcal{N}(2, \sqrt{3})$ <sup>6</sup>. Consideramos una m.a.s. de tamaño  $n=450$ . Con base en los resultados de las secciones anteriores, podemos suponer que la media muestral  $\bar{x}$  sigue una distribución normal  $\bar{x} \sim \mathcal{N}(2, \sqrt{\frac{3}{450}})$ . Como  $\mathbb{E}(\bar{x}) = 2$ , sabemos que  $\bar{x}$  debería estar cerca a 2 y, usando las propiedades de la distribución Normal, podemos calcular la probabilidad de que  $\bar{x}$  esté en cualquier intervalo centrado en 2; por ejemplo, para el intervalo  $[1, 90; 2, 10]$  la probabilidad es:

$$\mathbb{P}(1, 90 \leq \bar{x} \leq 2, 10) = 0, 78. \quad (3.4)$$

En otras palabras, el intervalo  $[1, 90; 2, 10]$  contiene a  $\bar{x}$  con un **nivel de confianza** de 78 %. Si pudiéramos hacer muchas muestras distintas de tamaño 450, esperaríamos que en el 78 % de ellas la media muestral se encuentre en este intervalo.

La Fórmula 3.4 puede también escribirse como:

$$\mathbb{P}(1, 90 \leq \bar{x} \leq 2, 10) = \mathbb{P}(\mu - 0, 10 \leq \bar{x} \leq \mu + 0, 10) = \mathbb{P}(\bar{x} - 0, 10 \leq \mu \leq \bar{x} + 0, 10), \quad (3.5)$$

lo que permite interpretar  $\mathbb{P}(\bar{x} - 0, 10 \leq \mu \leq \bar{x} + 0, 10)$  como la probabilidad de que la media poblacional desconocida  $\mu$  se encuentra en el intervalo  $[\bar{x} - 0, 10; \bar{x} + 0, 10]$ .

La Figura 3.8 muestra los intervalos de confianza del 78 % (a) y del de 95 % (b), para 50 muestras aleatorias simuladas con la distribución  $X \sim \mathcal{N}(2, \sqrt{3})$ . Los puntos negros muestran los intervalos que no contienen el 2, que en este caso, sabemos que es el verdadero valor de  $\mu$ . Hay más puntos negros a la izquierda que a la derecha. De hecho, si hiciéramos mucho más que 50 experimentos, esperaríamos ver un 78 % de puntos negros a la derecha y un 95 % a la izquierda.

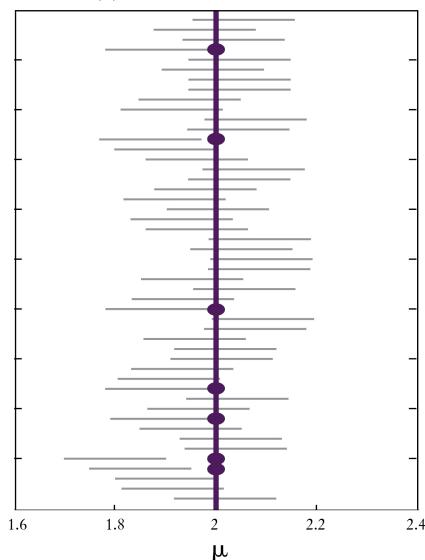
La Figura 3.9 ilustra la situación de otra manera.

---

<sup>6</sup>Este supuesto lo hacemos sólo por razones pedagógicas. Es obvio que en la práctica, si conocieramos la distribución de  $X$ , conoceríamos también su media y sería ocioso tratar de estimarla con una muestra.

FIGURA 3.8. Intervalos de confianza de 50 muestras

(a) Nivel de confianza 78%



(b) Nivel de confianza 95%

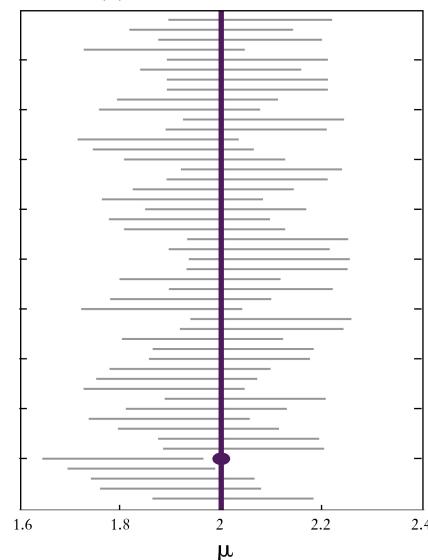
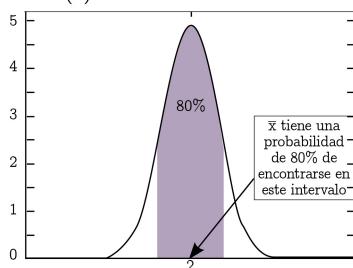
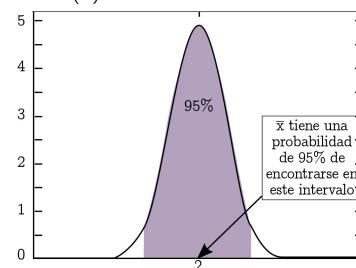


FIGURA 3.9. Niveles de confianza

(a) Nivel de confianza 80%



(b) Nivel de confianza 95%



¿Cómo calculamos el 78 % de la fórmula 3.4? Con un software estadístico o Excel, se pueden obtener directamente las probabilidades  $\mathbb{P}(v \leq 2, 10)$ , que vale 89 % y  $\mathbb{P}(\bar{x} \leq 1, 90)$  que vale 11 %, la diferencia da el nivel de confianza de 78 %. Cuando no se tiene software, se usa la tabla de distribución  $\mathcal{N}(0, 1)$  después de transformar la variable  $\bar{x}$  (Ver en Anexo la Tabla). Considerando que  $\bar{x} \sim \mathcal{N}(2, \sqrt{\frac{3}{450}})$ , se transforma la variable  $\bar{x}$  en una variable  $Z \sim \mathcal{N}(0, 1)$ :  $Z = \frac{\bar{x}-2}{\sqrt{3/450}}$ . Aplicando la transformación

en los dos términos de la probabilidad, se tiene, entonces,  $\mathbb{P}(\bar{x} \leq 1,90) = \mathbb{P}(Z \leq \frac{1,90-2}{\sqrt{3/450}}) = \mathbb{P}(Z \leq -1,22)$  y  $\mathbb{P}(\bar{x} \geq 2,10) = \mathbb{P}(Z \geq \frac{2,10-2}{\sqrt{3/450}}) = \mathbb{P}(Z \geq 1,22)$ . La tabla de la distribución  $\mathcal{N}(0, 1)$  da  $\mathbb{P}(Z \geq 1,22)$  que vale 0,11; por simetría se obtiene el mismo valor para  $\mathbb{P}(Z \leq -1,22)$  (Figura 3.9(a)).

De manera más general, si  $[a, b]$  es un intervalo de confianza para  $\mu$ , y  $c = \frac{b-a}{2}$ , entonces  $a = \bar{x} - c$  y  $b = \bar{x} + c$  y el ancho del intervalo es igual a  $2c$ . El problema es encontrar el valor  $c$  para un nivel de confianza dado.

Como en la Fórmula 3.5, podemos escribir:

$$\mathbb{P}(a \leq \mu \leq b) = \mathbb{P}(\bar{x} - c \leq \mu \leq \bar{x} + c) = \mathbb{P}(\mu - c \leq \bar{x} \leq \mu + c).$$

Como  $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ , su función de densidad es simétrica con respecto de  $\mu$ . Tenemos, entonces  $\mathbb{P}(\bar{x} \leq \mu - c) = \mathbb{P}(\bar{x} \geq \mu + c)$ . Sea  $\mathbb{P}(\bar{x} \leq \mu - c) = \mathbb{P}(\bar{x} \geq \mu + c) = \alpha$ , entonces  $\mathbb{P}(a \leq \mu \leq b) = 1 - \alpha$ . Se dice, entonces, que  $[a, b]$  es un intervalo de confianza para  $\mu$  de nivel de confianza  $1 - \alpha$ .

Si  $Z \sim \mathcal{N}(0, 1)$ ,

$$\mathbb{P}(\bar{x} \geq \mu + c) = \mathbb{P}\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{c}{\sigma/\sqrt{n}}\right) = \mathbb{P}(Z \geq \frac{c}{\sigma/\sqrt{n}}) = \alpha$$

Luego, se busca en la tabla  $N(0, 1)$  el valor  $u$  tal que  $\mathbb{P}(Z \geq u) = \alpha$ . Por ejemplo, para  $\alpha = 0,025$ ,  $u = 1,96$ .

Notemos que el intervalo de confianza representa la precisión de la media muestral. El ancho del intervalo  $I$ , que es igual a  $I = 2c = 2u\frac{\sigma}{\sqrt{n}}$ , depende de:

- El nivel de confianza  $1 - \alpha$  elegido. El ancho  $I$  crece cuando  $1 - \alpha$  crece.
- El tamaño  $n$  de la muestra. El ancho  $I$  decrece cuando  $n$  crece.
- La varianza  $\sigma^2$  de la población. El ancho  $I$  crece cuando  $\sigma^2$  crece.

En los ejemplos anteriores suponíamos conocidas la media y la varianza muestrales (2 y 3, respectivamente). En la práctica no conocemos ninguna de las dos cosas, pero podemos estimarlas a partir de la misma muestra. En una segunda monografía veremos las consecuencias que eso tiene para la estimación por intervalos. Adelantándonos a eso, digamos que, en el caso general, tendremos que recurrir a una nueva distribución de probabilidad: la distribución de Student. Sin embargo, cuando las muestras son suficientemente grandes, podemos simplemente reemplazar la media y la varianza por sus estimaciones muestrales, y seguir usando la distribución normal como lo hemos hecho aquí.

**Ejemplo 3.1.** El Sernac está inspeccionando las bolsas de azúcar producidas por la firma Acme, que supuestamente traen un kilo cada una. Se toma una muestra aleatoria simple de 850 bolsas de una partida. El peso promedio resulta ser 996 g, con una desviación típica de 50 g. Construyamos intervalos de confianza para el peso promedio real  $\mu$  de las bolsas, con los niveles de confianza de 90 %, 95 % y 99 %:

$1 - \alpha$	$\alpha/2$	$u$	Intervalo	Ancho del intervalo
90 %	5,0 %	1,645	[993,18; 998,82]	5,64
95 %	2,5 %	1,960	[992,64; 999,36]	6,72
99 %	0,5 %	2,576	[991,58; 1000,42]	8,84

Compare los intervalos de confianza en función del nivel de confianza. ¿En qué casos el intervalo contiene el valor anunciado de 1000g?, y ¿qué pasaría si la desviación típica de la muestra fuera de 60g?

### 3.5.2 Proporción

Consideramos ahora una variable  $X$  binaria (0-1) (por ejemplo, Cara/Sello; respuesta Sí/No a una pregunta; elección con dos candidatos).  $X$  sigue una distribución de Bernoulli,  $X \sim B(p)$ , donde el parámetro  $p = \mathbb{P}(X = 1)$  es la proporción (o **prevalencia**) de una de las alternativas en la población.

Como la proporción no es más que un caso particular de la media (¿por qué?) todos los conceptos desarrollados para la estimación de medias se aplican directamente a la estimación de proporciones, con una simplificación importante, dada por el hecho que en una distribución de Bernoulli no es necesario estimar la varianza separadamente de la media. En efecto, si conocemos la prevalencia  $p$ , conocemos automáticamente la varianza, que es  $p(1 - p)$ .

Para encontrar intervalos de confianza para  $p$ , procedemos de manera similar al caso de la media. El estimador de  $p$  es  $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$  (la proporción de valores  $X = 1$  en la muestra), que tiene una distribución aproximada  $\mathcal{N}\left(\hat{p}, \sqrt{\frac{p(1-p)}{n}}\right)$ .

Se obtiene así cómo el intervalo de confianza para  $p$  con nivel de confianza 95 %:

$$[\hat{p} - 1,96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1,96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

Ejemplo: El Ministerio de Educación hizo una encuesta para estimar la prevalencia  $p$  de fumadores entre los escolares chilenos. En una muestra aleatoria simple de 500 alumnos de cuarto medio se encontró un 32 % de fumadores. Construyamos los intervalos de confianza para  $p$  con los niveles de confianza 90 %, 95 % y 99 %.

Hay que buscar, en la Tabla de la distribución Normal,  $u$  tal que:  $\mathbb{P}(Z \geq u) = \alpha/2$  para un nivel de confianza  $1 - \alpha$ . Compare los intervalos que se encuentran en la tabla adjunta.

$1 - \alpha$	$\alpha/2$	$u$	Intervalo	Ancho del intervalo
90 %	5,0 %	1,645	[28,57; 35,43]	6,86
95 %	2,5 %	1,960	[27,91; 36,09]	8,18
99 %	0,5 %	2,576	[26,63; 37,37]	10,75

### 3.6 Tests de hipótesis

#### 3.6.1 Introducción

Muchos estudios estadísticos se plantean los tipos de preguntas siguientes:

- ¿Tiene el cambio en la malla curricular de la Enseñanza Media un impacto positivo sobre el aprendizaje de los alumnos?
- ¿Más del 25 % de los consumidores chilenos no miran los precios en el supermercado?
- ¿Las niñas chilenas son más altas que las japonesas?
- ¿Las niñas tienen mejor rendimiento escolar que los niños en Chile?
- ¿Un nuevo tratamiento médico es más efectivo que los antiguos?
- ¿El Citroen consume menos bencina que el Toyota?
- ¿Los chilenos comen más pan que los argentinos?
- ¿La próxima elección presidencial terminará en un empate técnico?
- ¿Es eficaz el régimen para adelgazar del Dr Gordillo?

Las respuestas dependen generalmente de datos recogidos sobre una muestra.

En 1935, Sir Ronald Fisher en su obra “Experimental Design” expuso el caso de una dama que afirmaba que, cuando probaba una taza de té con leche, podía distinguir si se había puesto primero el té o la leche. Algunas personas sospechaban que la dama respondía al azar y acertaba, a veces, por pura suerte.

Al igual que en el problema de estimación, nunca podremos determinar con absoluta certeza si la dama responde al azar, pero podemos estimar la probabilidad de que lo haga. Fisher propuso preparar ocho tazas, poniendo la leche antes del té en cuatro de ellas, y después del té, en las otras cuatro; luego las presenta al azar a la dama, para que las pruebe todas y diga qué se puso primero en cada una, la leche o el té. Si acierta en las ocho tazas, decidimos que efectivamente puede distinguirlas. Justificaba esta decisión diciendo que si la dama contesta al azar, sus chances de acertar en las ocho tazas son sólo 1 en 70, pues hay  $\binom{8}{4} = 70$  casos posibles, y uno sólo de ellos es favorable. En otras palabras, la probabilidad de que la dama esté contestando al azar, si acierta en las ocho tazas es  $\frac{1}{70}$  (Tabla 3.4). Por lo tanto, si decidimos que no contestó al azar, nuestra probabilidad de equivocarnos es muy pequeña:  $\frac{1}{70} = 0,0143$ . Esta probabilidad se llama **p-valor**.

La tabla siguiente muestra los resultados posibles del experimento de Fisher, expresados en términos de la cantidad de tazas “con leche antes del té” que la dama identifique.

TABLA 3.4. Resultados del experimento de Fisher

Resultados posibles	Número de casos favorables	Probabilidad si respondió al azar
4 correctas + 0 incorrecta	1	1/70
3 correctas + 1 incorrecta	16	16/70
2 correctas + 2 incorrectas	36	36/70
1 correcta + 3 incorrectas	16	16/70
0 correctas + 4 incorrectas	1	1/70

En el problema de la dama probando té con leche hay dos errores posibles:

1. Declarar que la dama sabe distinguir entre las tazas, cuando en realidad contestó al azar.
2. Declarar que la dama contestó al azar, cuando en realidad sabe distinguir entre las tazas.

Los investigadores científicos, los médicos y nosotros mismos en la vida cotidiana, nos enfrentamos frecuentemente con problemas similares. Tenemos que elegir entre dos hipótesis, y establecemos una regla para elegir una de ellas. Una de las hipótesis se llama **Hipótesis Nula**  $H_0$  y la otra, **Hipótesis Alternativa**  $H_1$ . La regla de decisión se llama **test estadístico**.

En el test de hipótesis, hay dos tipos posibles de error:

- Error de Tipo I: declarar  $H_1$  cierta cuando en realidad  $H_0$  es cierta.
- Error de Tipo II: declarar  $H_0$  cierta cuando en realidad  $H_1$  es cierta.

Las probabilidades de los errores Tipo I y Tipo II se designan convencionalmente con las letras griegas  $\alpha$  y  $\beta$ , respectivamente.

En toda toma de decisión, es importante evaluar las consecuencias de equivocarse. En la mayoría de los casos, los roles de las hipótesis no son los mismos y las consecuencias de equivocarse en un sentido o en otro no tienen la misma importancia. Habitualmente, la hipótesis nula es la más conservadora, la que representa la tradición, el consenso, la ciencia bien establecida; mientras que la hipótesis alternativa representa algo interesante, innovador o novedoso. Generalmente, se prefiere controlar la probabilidad  $\alpha$  de cometer un error de Tipo I, tratando de que la probabilidad  $\beta$  de cometer un error de Tipo II sea lo más pequeña posible.

La situación es análoga a la de un juez, que debe declarar a un acusado culpable o inocente, en base a las evidencias presentadas por la policía.

El juez tiene dos formas de equivocarse: dejar libre a un culpable o condenar a un inocente. En la mayoría de los países, el derecho no les da a los dos errores la misma importancia: se considera que es más grave condenar a un inocente que liberar a un culpable. Basándose en este principio, el juez va a tratar de controlar el error de “condenar un sospechoso que podría ser inocente” y, si las evidencias no le permiten establecer en forma fehaciente la culpabilidad del sospechoso, va a preferir dejarlo libre. (Es interesante recordar que la jurisprudencia no ha sido siempre la misma. En 1209 el enviado del Papa Inocencio III recomendó matar a todos los habitantes de la ciudad francesa de Béziers, pues sabía que muchos eran herejes. Cuando le recordaron que entre los ciudadanos también había fieles, dijo “No importa, Dios los reconocerá.”)

Ejemplos de hipótesis (que podrían ser nulas o alternativas dependiendo del interés del investigador) podrían ser:

- H: la media de la población es igual a 1.
- H: la media de la población es  $\leq 1$ .
- H: la probabilidad de sacar “cara” al lanzar una moneda es mayor que 0,5.
- H: la varianza de la población es  $\geq 2$ .
- H: la distribución de población es normal.

Las cuatro primeras hipótesis se refieren a un parámetro de la distribución de población, mientras que la última se refiere al tipo de distribución. En el primer caso se habla de **hipótesis paramétricas** y en el segundo, de **hipótesis no paramétricas**. En esta sección nos ocuparemos de hipótesis paramétricas relativas a medias y a proporciones. Veamos la construcción de una regla de decisión con un ejemplo.

En el proceso de fabricación de una bebida, se sabe que la producción por cadena tiene una media diaria de 500 litros, con una desviación estándar de 100 litros. Se propone una modificación del proceso, con el objeto de aumentar la producción diaria. Se implementó el nuevo proceso sobre una de la cadena, que en una muestra de 60 días dio una producción promedio de 525 litros, ¿podemos decir que el nuevo proceso es eficaz?

Sea  $\mu_o = 500$  la producción de proceso de fabricación antiguo y  $\mu$  la producción actual. Queremos decidir entre las hipótesis:  $H_o : \mu = \mu_o$  y  $H_1 : \mu > \mu_o$ .

El p-valor del test es la probabilidad de que rechacemos  $H_o$  por haber encontrado en la muestra un valor consistente con la hipótesis alternativa (525 litros, específicamente), cuando en realidad  $H_o$  era cierta. Para calcularlo, recordemos que  $\bar{x}$  tiene



una distribución aproximadamente normal:  $\bar{x} \sim \mathcal{N}(\mu, \sigma/\sqrt{60})$  con  $\sigma = 100$ . Hagámoslo bajo las hipótesis  $H_0: \mu = \mu_o = 500$  y  $H_1: \mu > \mu_o$ .

Si vamos a rechazar  $H_0$  porque  $\bar{x}$  resultó ser 525 litros, con mayor razón lo habríamos hecho si  $\bar{x}$  hubiera resultado ser mayor que 525 litros. El p-valor es, entonces, la probabilidad de que  $\bar{x}$  sea mayor o igual a 525 litros, siendo que en la población la media  $\mu = \mu_o = 500$ . Esta probabilidad se calcula con  $\bar{x} \sim \mathcal{N}(500, 100/\sqrt{60})$ , o sea, con la normal  $Z \sim \mathcal{N}(0, 1)$ :  $Z = \frac{\bar{x} - 500}{100/\sqrt{60}}$ :

$$p\text{-valor} = \mathbb{P}(\bar{x} \geq 525) = \mathbb{P}\left(Z \geq \frac{525 - 500}{100/\sqrt{60}}\right) = \mathbb{P}(Z \geq 1, 936) = 0, 0264.$$

Como el p-valor es relativamente pequeño, podemos decir que hay bastante evidencia de que el nuevo proceso permite aumentar la producción diaria de bebida, y recomendar que se implemente el nuevo proceso sobre las otras cadenas de producción. El p-valor es la probabilidad de cometer un error de Tipo I al declarar que  $H_1$  es cierta si en la muestra  $\bar{x}$  es al menos igual a 525 litros. También, podríamos haber fijado la probabilidad de cometer un error de Tipo I (por ejemplo, decir que queremos que  $\alpha$  sea igual a 1%), y buscar los valores de  $\bar{x}$  que nos inducirían a incurrir en ese error:  $\mathbb{P}(\bar{x} \geq a) = 0, 01$ , donde  $a$  es el valor mínimo de  $\bar{x}$  para tener un  $\alpha$  de 1% y  $\bar{x} \sim \mathcal{N}(500, 100/\sqrt{60})$ . Ese valor de  $a$  se encuentra haciendo:

$$\mathbb{P}(\bar{x} \geq a) = \mathbb{P}\left(Z \geq \frac{a - 500}{100/\sqrt{60}}\right) = 0, 01$$

En la tabla de la distribución Normal  $\mathcal{N}(0, 1)$ , obtenemos:

$$\mathbb{P}(Z \geq 2, 326) = 0, 01$$

Luego,

$$\frac{a - 500}{100/\sqrt{60}} = 2, 326 \implies a = 530, 03$$

Si quisiéramos que  $\alpha$  fuera de sólo 1%, no podríamos concluir que el nuevo proceso es más eficaz que el antiguo. Para hacerlo, necesitaríamos que  $\bar{x}$  fuera mayor que 530,03 litros, pero en la muestra sólo obtuvimos 525.

Se dice que el conjunto  $\mathcal{R} = \{\bar{x} \geq 530, 03\}$  es la **región crítica** correspondiente a  $\alpha = 1\%$ .

Compruebe que en la región crítica correspondiente a un valor menos exigente,  $\alpha = 5\%$  es igual a  $\{\bar{x} \geq 521, 24\}$ , y que en este caso aceptaríamos el nuevo proceso. La decisión depende del riesgo que estamos dispuesto a asumir.

Ahora, ¿qué pasa con  $\beta$ , la probabilidad del error de tipo II? La hipótesis  $H_1$  no establece un valor específico de la media  $\mu$  – sólo dice que es mayor que  $\mu_o$ , pero hay infinitos valores de  $\mu$  mayores que  $\mu_o$ . Podemos calcular  $\beta$  para cada hipótesis alternativa posible  $H_1$ , manteniendo constante  $\alpha$  en 1%, recordando que  $\beta$  es la probabilidad de no rechazar  $H_0$ , o sea, la probabilidad de que encontramos  $\bar{x} < 530, 03$

cuando  $\mu = 520$ . Por ejemplo, para  $\mu = 520$ , tenemos que:

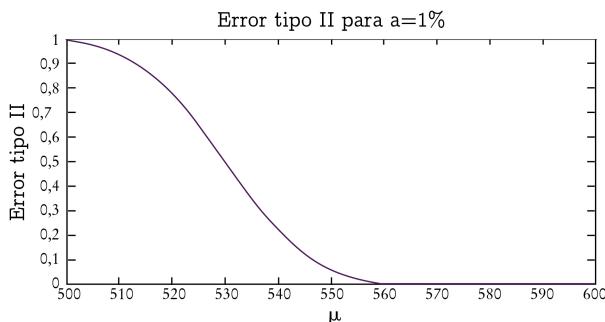
$$\beta = \mathbb{P}(\bar{x} < 530,03) = \mathbb{P}(Z < \frac{530,03 - 520}{100/\sqrt{60}}) = \mathbb{P}(Z < 0,777) = 0,781$$

Este error es bastante grande. La tabla 3.5 y la Figura 3.10 muestran  $\beta$  para valores de  $\mu$  entre 500 y 570 litros;  $\beta$  disminuye a medida que  $\mu$  se aleja de  $\mu_o = 500$ , lo que no debería sorprendernos, pues es obviamente más fácil decidir entre valores muy distintos que entre valores parecidos.

TABLA 3.5. Error de Tipo II

$\mu$	500	510	520	530	540	550	560	570
Error de Tipo II $\beta$	0,990	0,940	0,781	0,501	0,220	0,061	0,010	0,001

FIGURA 3.10. Error de Tipo II

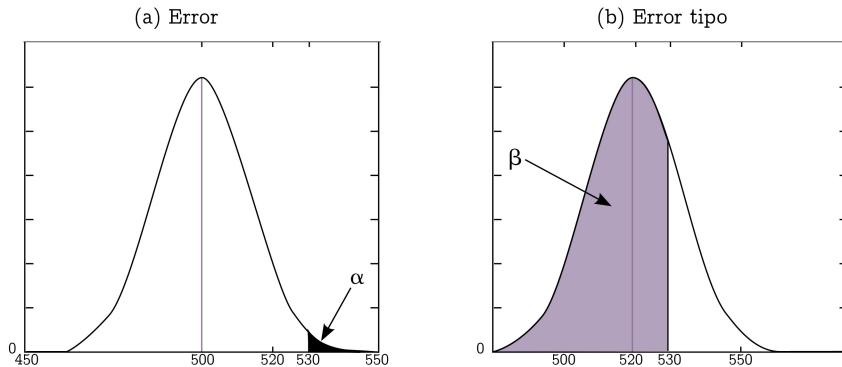


La probabilidad  $1 - \beta$  de no equivocarse cuando  $H_1$  es cierta se llama **potencia**. Minimizar  $\beta$  es equivalente a maximizar la potencia.

Si quisiéramos disminuir  $\beta$  para  $\mu = 520$  (que vale 0,781 cuando  $\alpha$  vale 1 %), hay que aumentar  $\alpha$ . La Figura 3.11 lo muestra gráficamente: para disminuir  $\beta$  (el área morada del gráfico (b)), hay que correr el valor límite (530,03) hacia la izquierda, pero esto hace aumentar  $\alpha$  (el área negra del gráfico (a)).

Los conceptos de p-valor y región crítica son complementarios: el p-valor da la probabilidad mínima de equivocarnos si rechazamos la hipótesis  $H_o$  con el valor encontrado en la muestra; la región crítica es el conjunto de valores de  $\bar{x}$  para los cuales se rechaza  $H_o$  con un error controlado. En ambos casos, la regla de decisión depende del riesgo que estemos dispuestos a asumir. La probabilidad  $\alpha$  de cometer un error de Tipo I se denomina también **nivel de significación**. Es común referirse a los errores de Tipo I como “falsos positivos” y a los de tipo II como “falsos negativos”.

FIGURA 3.11. Errores de Tipo I y II



### 3.6.2 Tests para la media de una población

Aquí desarrollamos la teoría de test de hipótesis relativa a una media, utilizando la distribución normal. Consideramos tres casos:

#### Caso 1

El Ministerio de Salud se pregunta si las niñas chilenas son más altas que las de las anteriores generaciones. En un estudio realizado en 1950 se midieron todas las niñas de 10 años y se registró una estatura promedio de 130 cm, con una desviación estándar de 7 cm. En una m.a.s. de 200 niñas de la misma edad medidas este año, se obtuvo un promedio  $\bar{x}$  de 132 cm. La pregunta es si acaso las niñas actuales son más altas que sus antepasadas, o si no ha habido cambio y la diferencia de 2 cm se debe tan sólo al error muestral.

Sea  $\mu$  la estatura media de todas las niñas de 10 años de la actual generación. (Esa media es desconocida; sólo conocemos la media de una muestra.) Queremos decidir entre las hipótesis:  $H_0 : \mu = 130\text{cm}$  y  $H_1 : \mu > 130\text{cm}$ .

El p-valor del test es la probabilidad de que rechacemos  $H_0$  por haber encontrado en la muestra un valor 132 cm, cuando en realidad  $H_0$  era cierta. Para calcularlo, recordemos que  $\bar{x}$  tiene una distribución aproximadamente normal:  $\bar{x} \sim \mathcal{N}(\mu, \sigma/\sqrt{200})$  con  $\sigma = 7$ .

Como en el ejemplo anterior, si vamos a rechazar  $H_0$  porque  $\bar{x}$  resultó ser 132 cm, con mayor razón lo habríamos hecho si  $\bar{x}$  hubiera resultado ser mayor que 132 cm. El p-valor es, entonces, la probabilidad de que  $\bar{x}$  sea mayor o igual a 132 cm, dado que en la población la media  $\mu = \mu_0 = 130\text{cm}$ . Esta probabilidad se calcula con  $\bar{x} \sim \mathcal{N}(130, 7/\sqrt{200})$ , o sea con la normal  $Z \sim \mathcal{N}(0, 1)$ :  $Z = \frac{\bar{x} - 130}{7/\sqrt{200}}$ :

$$p\text{-valor} = \mathbb{P}(\bar{x} \geq 132) = \mathbb{P}\left(Z \geq \frac{132 - 130}{7/\sqrt{200}}\right) = \mathbb{P}(Z \geq 4,041) = 0,000027$$

Como el p-valor es muy pequeño, podemos decir que hay bastante evidencia de que las niñas de 10 años son más altas ahora que antes.

El p-valor es la probabilidad de cometer un error de Tipo I al declarar que  $H_1$  es cierta, si en la muestra  $\bar{x}$  es, al menos, igual a 132 cm. Fijemos ahora la probabilidad de cometer un error de Tipo I en el valor  $\alpha = 1\%$ , y busquemos los valores de  $\bar{x}$  que nos inducirían a cometer ese error: el menor de esos valores sería el número  $a$  tal que  $\mathbb{P}(\bar{x} \geq a) = 0,01$ , si  $\bar{x} \sim \mathcal{N}(130, 7/\sqrt{200})$ , y se puede encontrar haciendo:

$$\mathbb{P}(\bar{x} \geq a) = \mathbb{P}(Z \geq \frac{a - 130}{7/\sqrt{200}}) = 0,01$$

En la tabla de la distribución Normal  $\mathcal{N}(0, 1)$ , obtenemos:

$$\mathbb{P}(Z \geq 2,326) = 0,01$$

Luego,

$$\frac{a - 130}{7/\sqrt{200}} = 2,326 \implies a = 131,15$$

La región crítica correspondiente a  $\alpha = 1\%$  es  $\{\mathcal{R} = \bar{x} \geq 131,15\}$ . Como el valor de la muestra (132 cm) pertenece a esa región crítica, aceptamos que las niñas actuales son más altas que sus antepasadas con un riesgo de 1%. Obviamente, lo habríamos aceptado también si hubiéramos trabajado con  $\alpha = 5\%$ , o con  $\alpha = 0,5\%$ , o con cualquier valor de  $\alpha$  mayor que el p-valor, que es prácticamente nulo en este caso.

Ahora consideremos el error de Tipo II. Calculemos  $\beta$  para cada valor posible de  $H_1$ , manteniendo constante  $\alpha$  en 1%.

Por ejemplo, si  $\mu = 131$ ,  $\beta$  es la probabilidad de que  $\bar{x}$  sea menor que 131,15, dado que  $\mu = 131$  (Figura 3.12(a)):

$$\beta = \mathbb{P}(\bar{x} < 131,15) = \mathbb{P}(Z < \frac{131,15 - 131}{7/\sqrt{200}}) = \mathbb{P}(Z < 0,30) = 0,619$$

La tabla 3.6 muestra cómo disminuye  $\beta$  para valores de  $\mu$  entre 130 cm y 132,5 cm.

TABLA 3.6. Error de Tipo II n=200

$\mu$	130,0	130,5	131,0	131,5	132,0	132,5
Error de Tipo II $\beta$	0,990	0,806	0,620	0,241	0,043	0,003

¿Habrían sido distintas nuestras conclusiones si los resultados hubieran provenido de una muestra de 500 niñas? El p-valor sería:

$$p\text{-valor} = \mathbb{P}(\bar{x} \geq 132) = \mathbb{P}(Z \geq \frac{132 - 130}{7/\sqrt{500}}) = \mathbb{P}(Z \geq 6,389) = 0,000000000083$$

La región crítica sería  $\{\bar{x} \geq 130,73\}$ , para un  $\alpha$  de 1 %, y los valores de  $\beta$  (Tabla 3.7) son aún más pequeños que antes para los mismos valores de la hipótesis alternativa. Es mucho más “fácil” rechazar la hipótesis nula cuando la muestra es más grande, pues lo hacemos a partir de un valor más cercano a  $\mu_o$ .

TABLA 3.7. Error de Tipo II, n=500

$\mu$	130,0	130,5	131,0	131,5	132,0	132,5
Error de Tipo II $\beta$	0,990	0,769	0,194	0,007	0,000	0,000

### Caso 2

Una automotriz lanza al mercado un vehículo, afirmando que rinde en promedio 10 km por litro. Un taller decide verificarlo, midiendo el rendimiento de 75 vehículos del mismo modelo. Obtiene un promedio de 9,8 km/l con una desviación estándar de 0,9 km/l. Consideramos las hipótesis  $H_o : \mu = 10$  y  $H_1 : \mu < 10$ .

La región crítica para rechazar hipótesis nula es  $\mathcal{R} = \{\bar{x} \leq a\}$ , donde  $a$  es tal que  $\mathbb{P}(\bar{x} \leq a) = \alpha$ . Para  $\alpha = 1\%$ ,  $\mathbb{P}(Z \leq 2,326) = \mathbb{P}(\bar{x} \leq 10 - 2,326 \frac{\sqrt{75}}{9,8}) = 1\%$ . O sea,  $\mathcal{R} = \{\bar{x} \leq 9,758\}$ . Como en la muestra encontramos una media igual a 9,8 km/l, no se rechaza la hipótesis nula con  $\alpha = 1\%$ . Si hubiéramos tomado  $\alpha = 5\%$ , la región crítica hubiera sido  $\mathcal{R} = \{\bar{x} \leq 9,829\}$ , lo nos habría llevado a la conclusión contraria. Calculemos el p-valor:

$$p\text{-valor} = \mathbb{P}(\bar{x} \leq 9,8) = \mathbb{P}(Z \leq \frac{9,8 - 10}{0,9/\sqrt{75}}) = \mathbb{P}(Z \geq -1,924) = 0,027$$

Con esta muestra, concluiríamos que el vehículo tiene un rendimiento inferior a 10 km/l con cualquier  $\alpha$  mayor que 2,7 %.

### Caso 3

Un hospital recibe una gran partida de bolsas de solución de suero fisiológico, que supuestamente deberían contener 50 mg de suero cada una. Una muestra aleatoria de 80 bolsas dio una media de 49,2 mg, con una desviación estándar de 2 mg, ¿el hospital debería aceptar los frascos?

Si  $\mu$  es la media de todas las bolsas de la partida, las hipótesis a plantear son:  $H_o : \mu = 50$  contra  $H_1 : \mu \neq 50$ . En éste la hipótesis alternativa no tiene como antes un signo “>” o “<”, sino un signo “≠”, porque idealmente las bolsas deberían tener exactamente 50 mg de suero, ni más ni menos. Eso significa que la región crítica va a estar constituida por dos regiones, una de cada lado del valor 50:

$$\mathcal{R} = \{\bar{x} \leq a_1\} \cup \{\bar{x} \geq a_2\}$$

Las hipótesis de este tipo se llaman “bilaterales”; las que vimos anteriormente, en los casos 1 y 2, se llaman “unilaterales”.

La región crítica para un  $\alpha$  dado, se calcula repartiéndolo en dos partes iguales:  $\alpha/2$  a la izquierda y  $\alpha/2$  a la derecha:

$$\mathbb{P}(\bar{x} \leq a_1) = \mathbb{P}(\bar{x} \geq a_2) = \alpha/2$$

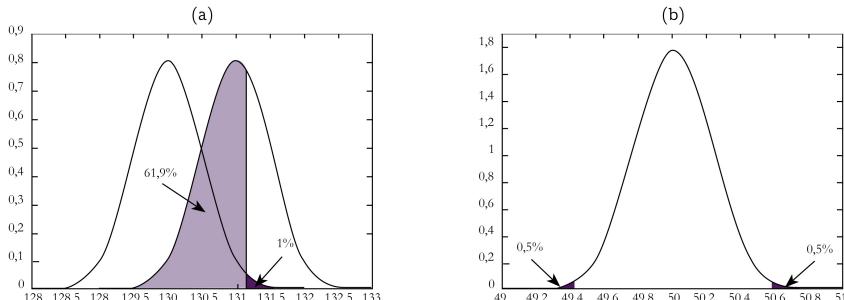
Para  $\alpha = 1\%$ , bajo  $H_o$ ,  $\bar{x} \sim \mathcal{N}(50, 2/\sqrt{80})$ , obtenemos:

$$\mathbb{P}(Z \leq -2,576) = 0,005 \implies \mathbb{P}(\bar{x} \leq 50 - 2,576 \frac{2}{\sqrt{80}}) = \mathbb{P}(\bar{x} \leq 49,42) = 0,005$$

$$\mathbb{P}(Z \geq 2,576) = 0,005 \implies \mathbb{P}(\bar{x} \geq 50 + 2,576 \frac{2}{\sqrt{80}}) = \mathbb{P}(\bar{x} \geq 50,58) = 0,005$$

Para no rechazar  $H_o$  con  $\alpha = 1\%$ , la media de las 80 bolsas debería estar entre 49,42 y 50,58. Como fue 49,2, el hospital debería devolver la partida de suero (Figura 3.12(b)).

FIGURA 3.12. Región crítica del test



En resumen, para tomar una decisión entre dos alternativas a partir de datos muestrales, los pasos son:

1. Formular la hipótesis nula y la hipótesis alternativa en función de lo que quiere poner en evidencia.
2. Elegir un nivel de significación (probabilidad de cometer un error de tipo I)  $\alpha$ .
3. Determinar un estadístico y su distribución bajo la hipótesis nula.
4. Determinar la región crítica correspondiente al nivel de significación elegido.
5. Deducir la regla de decisión.
6. Calcular el p-valor.
7. Tomar la decisión: Si el p-valor es muy pequeño, se rechaza la hipótesis nula. Si se sospechaba que la hipótesis nula era falsa, pero el p-valor no resultó tan pequeño como para rechazarla, tal vez habría que tomar una muestra más grande para confirmar la convicción.

### 3.6.3 Test para la proporción de una población

Volvamos al ejemplo de la moneda cargada. Si sospechamos que la moneda está cargada hacia “cara”, planteamos hipótesis sobre el parámetro  $p$  de la variable de Bernoulli en donde  $p = \mathbb{P}(\text{cara})$ :  $H_0 : p = 0,5$  (moneda equilibrada) contra  $H_1 : p > 0,5$  (moneda cargada hacia “cara”).

Construyamos la región crítica  $\mathcal{R}$  para el número  $S$  de “caras” obtenido en  $n=120$  lanzamientos. Considerando la hipótesis  $H_1$ ,  $\mathcal{R}$  es de la forma  $\mathcal{R} = \{S \geq a\}$  de tal manera que  $\mathbb{P}(S \geq a) = \alpha$ .

El indicador  $S \sim \text{Binomial}(120; 0,5)$ . Si tomamos  $\alpha = 5\%$ ,  $\mathbb{P}(S \geq 69) = 5\%$ . La región crítica es, entonces,  $\mathcal{R} = \{S \geq 69\}$ .

Si resultara muy difícil calcular la probabilidad a partir de la distribución Binomial, se puede usar la aproximación a la Normal:  $S \sim \mathcal{N}(60, \sqrt{120 \times 0,5 \times 0,5})$ , o sea, suponer que  $S \sim \mathcal{N}(60, \sqrt{30})$ . Entonces:

$$\mathbb{P}(Z \geq 1,645) = 5\% \implies \mathbb{P}(S \geq 60 + 1,645 \times \sqrt{30}) = \mathbb{P}(S \geq 69,01) = 5\%$$

El procedimiento aproximado lleva prácticamente a la solución exacta.

Los casos  $H_0 : p \geq p_o$ , contra  $H_1 : p < p_o$ , y  $H_0 : p = p_o$ , contra  $H_1 : p \neq p_o$  se resuelven como en la sección anterior. Se deja como ejercicio.

Veamos otro ejemplo. La Superintendencia de Telecomunicaciones (SUBTEL) encargó un estudio de cobertura de una compañía de teléfonos celulares. La cobertura debería ser como mínimo un 90 % del territorio que pretende cubrir la compañía. Se eligieron al azar 800 puntos del territorio y se intentó llamar desde cada uno de ellos, con éxito en 710 casos (un porcentaje de 88,75 %, ¿puede la SUBTEL multar la compañía?

Veamos primero las hipótesis a contrastar. La variable de interés es una variable binaria:  $X = 1$  si la llamada es exitosa y  $X = 0$  en caso contrario.  $X \sim \text{Bernoulli}(p)$ , donde  $p$  es la probabilidad de una llamada exitosa. Tenemos, entonces, la hipótesis nula  $H_0 : p \geq 0,90$ , contra  $H_1 : p < 0,90$ . El número de llamadas exitosas  $S$  sigue una distribución  $\text{Binomial}(800, p)$ .

La SUBTEL sólo debería multar a la compañía si está muy segura que ésta no cumple la cobertura ofrecida de 90 %, y necesita controlar el error de declarar que la compañía no cumple cuando en realidad cumple.

Construyamos la región crítica para el número  $S$  de llamadas exitosas, con  $\alpha = 5\%$  y las hipótesis  $H_0 : p \geq 0,90$ , contra  $H_1 : p < 0,90$ . Sabemos que  $S \sim \text{Binomial}(800, 0,90)$  y buscamos el valor de  $a$  tal que  $\mathbb{P}(S \leq a) = 0,05$ . Ese valor resulta ser  $a=706$ . Como pudimos hacer 710 llamadas exitosas, con  $\alpha=5\%$  y una muestra de 800 llamadas no se puede multar la compañía.

Este cálculo también se puede hacer con una aproximación a la normal:  $S \sim \mathcal{N}(720, \sqrt{72})$  donde  $720 = 800 \times 0,9$  y  $72 = 800 \times 0,9 \times (1 - 0,9)$ . Se obtiene prácticamente la misma región crítica.

### 3.6.4 Test de comparación de dos medias

Muchos problemas de decisión se basan en comparación de dos poblaciones o dos mediciones.

Por ejemplo, para comparar la altura de las mujeres chilenas con la de las japonesas, el test de hipótesis consiste en comparar las estaturas medias de las dos poblaciones.

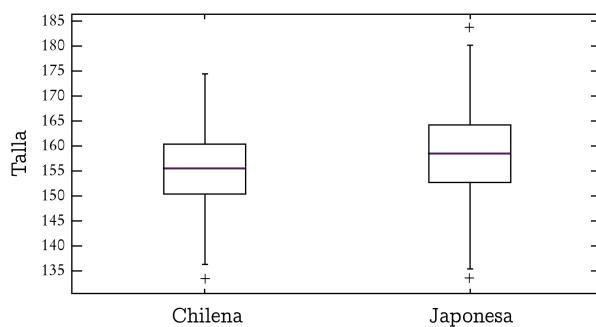
Un segundo ejemplo sería comprobar la eficacia de un régimen para adelgazar. El test de hipótesis consistiría en comparar dos medias en la misma población, antes y después del régimen.

En el primer caso los sujetos son diferentes y las dos muestras son independientes, mientras que en el segundo, c hay una sola muestra con dos mediciones por las cuales se quiere comparar las medias.

#### Comparación de medias en dos poblaciones

En una muestra de 350 mujeres chilenas se obtuvo una estatura media de 155 cm, mientras que en una muestra de 450 japonesas, el resultado fue 158 cm. ¿Podemos decir que las japonesas son más altas que las chilenas? Una primera mirada al diagrama de cajas no permite establecerlo con certeza (Figura 3.13).

FIGURA 3.13. Diagrama de cajas



Si  $\mu_1$  y  $\mu_2$  son las tallas medias de la talla de las chilenas y de las japonesas respectivamente, veamos las hipótesis nula y alternativa:  $H_0 : \mu_1 = \mu_2$  y  $H_1 : \mu_2 - \mu_1 > 0$ . Para construir el estadístico del test tenemos que considerar las distribuciones de las medias y desviaciones estándares en las dos muestras:  $\bar{x}_1$  y  $\sigma_1$  la media y desviación estándar de las tallas de las chilenas y  $\bar{x}_2$  y  $\sigma_2$  la media y desviación estándar de las tallas de las japonesas.

Estudios pasados permitieron tener desviaciones estándares de 7 cm para las chilenas y 7,5 cm para las japonesas. Estas desviaciones estándares, que no fueron obtenidas de las muestras, se consideran más fiables que las obtenidas de las mismas muestras.

Tenemos, de este modo,  $\bar{x}_1 \sim \mathcal{N}(\mu_1, \sigma_1/\sqrt{350})$  y  $\bar{x}_2 \sim \mathcal{N}(\mu_2, \sigma_2/\sqrt{450})$ . El indicador del test es, entonces, la diferencia  $\bar{x}_2 - \bar{x}_1$ . Como las muestras en Chile y Japón son independientes una de la otra, la varianza de  $\bar{x}_2 - \bar{x}_1$  es igual a la suma de las dos varianzas:  $\text{Var}(\bar{x}_2 - \bar{x}_1) = \text{Var}(\bar{x}_2) + \text{Var}(\bar{x}_1) = \frac{\sigma_1^2}{350} + \frac{\sigma_1^2}{450} = 0,265$ .

Se deduce la distribución del indicador:

$$\bar{x}_2 - \bar{x}_1 \sim \mathcal{N}(\mu_2 - \mu_1; 0,515),$$

donde  $0,515 = \sqrt{0,265}$

Bajo la hipótesis nula, la diferencia de las medias es nula y se obtiene la región crítica para  $\alpha=5\%$ :

$$\mathbb{P}(Z \geq 1,645) = \mathbb{P}(\bar{x}_2 - \bar{x}_1 \geq 1,645 \times 0,515) = \mathbb{P}(\bar{x}_2 - \bar{x}_1 \geq 0,847) = 0,05$$

La diferencia de las dos medias, que es igual a 3, se encuentra en la región crítica de 5%:  $\mathcal{R} = \{\bar{x}_1 - \bar{x}_2 \geq 0,847\}$ .

El p-valor resulta ser:

$$\mathbb{P}(\bar{x}_2 - \bar{x}_1 \geq 3) = \mathbb{P}\left(Z \geq \frac{3}{0,515}\right) = \mathbb{P}(Z \geq 5,825) = 0,000000029$$

lo que lleva a concluir que las japonesas son significativamente más altas de las chilenas.

### Comparación de dos medias en una población

Para estudiar si un régimen para adelgazar es eficaz, se hizo un estudio sobre una muestra aleatoria de 550 mujeres que siguieron el régimen durante seis meses. Las 550 mujeres se pesaron antes de empezar el régimen y después de seis meses. Tenemos dos mediciones relativas a las mismas mujeres: una media de 80 kg con una desviación estándar de 6,5 kg antes del régimen y de 74,6 kg con una desviación estándar de 10,5 kg después del régimen.

Sean  $\mu_1$  y  $\mu_2$  las medias antes y después del régimen. Quisiéramos decidir si  $\mu_2$  es significativamente menor que  $\mu_1$ , pero para eso no podemos usar el mismo método que en el ejemplo anterior, pues las muestras no son independientes. El del peso de una mujer después del régimen depende del que tenía antes. Tenemos que trabajar con las diferencias de los pesos individuales.

Sean  $\{x_1, x_2, \dots, x_n\}$  los pesos antes del régimen y  $\{y_1, y_2, \dots, y_n\}$  los pesos después. Calculamos, , entonces, las diferencias  $d_i = y_i - x_i$  y obtenemos el promedio de las 550 diferencias, que es igual a  $\hat{d} = -5,4$  kg con una desviación estándar de  $s = 7,96$  kg<sup>7</sup>. Sea  $\delta$  la media de la diferencia de pesos en la población de mujeres. Planteamos, entonces, las hipótesis nula y alternativa:  $H_0 : \delta = 0$  y  $H_1 : \delta < 0$ , y aplicamos el test de media del caso 2.

---

<sup>7</sup> $\hat{d} = 74,6 - 80 = -5,16$ , pero  $s^2 < 6,5^2 + 10,5^2$ , debido a la dependencia entre los valores de  $x_1$  y  $x_2$ .

La región crítica de nivel 5% es de la forma  $\mathcal{R} = \{\hat{d} \leq a\}$ , donde  $a$  se obtiene de la distribución de  $\hat{d} \sim \mathcal{N}(0; 7, 96/\sqrt{550})$ :

$$\mathbb{P}(Z \leq 1,645) = 0,05 \implies \mathbb{P}(\hat{d} \leq 1,645 \frac{\sqrt{550}}{7,96}) = 0,05 \implies a = -0,56$$

La región crítica para un error de tipo de 5% es  $\mathcal{R} = \{\hat{d} \leq -0,56\}$ . Eso permite concluir al éxito del régimen. Compruebe que el p-valor es prácticamente nulo y que, entonces, se puede rechazar  $H_o$  con un riesgo muy inferior a 5%.

### 3.7 Resumen de la terminología

Población o universo: El conjunto de todos los objetos que se quiere estudiar.

Unidad estadística: Elemento sobre lo cual se hacen mediciones.

Muestra: Subconjunto de la población.

Marco muestral: El marco muestral es el mecanismo que permite identificar los elementos de una población.

Parámetro: Característica de la población que se busca estimar.

Valores muestrales: Valores de variables obtenidos en la muestra.

Estimador: Una función de los valores muestrales.

Estimación: Valor del estimador obtenido de una muestra.

Muestreo: Método para obtener una muestra.

Muestreo aleatorio: Método de muestreo que selecciona la muestra de manera aleatoria.

Muestreo equiprobable: Muestreo que otorga la misma probabilidad de selección a todos las unidades de la población.

Muestreo aleatorio simple: Es un muestreo equiprobable en el cual se extraen las unidades una por una, de manera de que cada nueva unidad se obtiene del conjunto de las unidades no extraídas con equiprobabilidad.

Distribución en la población:

Distribución de los valores de la variable de interés.

Distribución en el muestreo:

Distribución de probabilidad de un estimador sobre todas las muestras posibles del mismo tamaño.

Error muestral: Error producido por el muestreo (método y tamaño de la muestra).

Errores no muestrales: Errores que provienen del proceso de mediciones sobre los elementos de la muestra.

Estimador insesgado: La media de la distribución muestral de un estimador es igual al parámetro de la población.

Desviación estándar: Es la raíz cuadrada de la varianza de una variable en la población.

Error estándar: Es la raíz cuadrada de la varianza de la distribución de un estimador obtenida a partir de todas las muestras del mismo tamaño.

Precisión: Es la medida de cuán cercano es un estimador del verdadero valor de un parámetro.

Nivel de confianza: Probabilidad asociada a un intervalo de confianza de un parámetro.

Error de Tipo I  $\alpha$ : Rechazar la hipótesis nula cuando ésta es cierta.

Error de Tipo II  $\beta$ : Aceptar la hipótesis nula cuando ésta es falsa.

Región crítica o región de rechazo: Es el conjunto de valores del estadístico en un test de hipótesis para los cuales la hipótesis nula es rechazada.

Potencia: Probabilidad que mide la habilidad de un test para rechazar la hipótesis nula cuando ésta es falsa. Es la

probabilidad de tomar una decisión correcta. Vale  $1 - \beta$ .

Nivel de significación: Probabilidad del error de tipo I que está dispuesto a asumir.

p-valor: Probabilidad de obtener el valor del estadístico del test tan extremo o más cuando la hipótesis nula es cierta.

### 3.8 Ejercicios

Los ejercicios con \* pueden utilizarse con los estudiantes de Enseñanza Media.

1. Se confecciona el marco muestral y se identifican los elementos de la población en estudio con números correlativos desde el 1 hasta  $N$ , donde  $N$  es el último elemento de la población. Luego, se seleccionan números al azar y se encuesta a las personas correspondientes con dicho número. Lo anterior corresponde a: (a) Un muestreo estratificado; (b) Un muestreo aleatorio simple; (c) Un muestreo sistemático.

2. (\*) En una escuela hay 1200 estudiantes (niños y niñas). Si en una muestra de 100 estudiantes elegidos al azar hay 45 niños, ¿cuál de las siguientes alternativas es el número más probable de niños en la escuela?<sup>8</sup> 450, 500, 540 ó 600.

3. (\*) De un lote de 3000 ampolletas, 100 fueron seleccionadas al azar y probadas. Si 5 de las ampolletas de la muestra estaban quemadas, ¿aproximadamente cuántas ampolletas quemadas se esperaría encontrar en el lote completo?<sup>9</sup>

4. Una muestra estará correctamente sacada cuando: (I) Sea aleatoria; (II) Sea sesgada; (III) Sea a partir de toda la población; (IV) Posea un tamaño adecuado.

¿Cuál alternativa es correcta? (a) I y II; (b) I y III; (c) II y IV; (d) I, III, IV; (e) Todas

5. Pedro juega los números 1, 2, 3, 4, 5 y 6 al loto. Juan le dice que tiene menos posibilidad de ganar con esta serie que jugando seis números al azar. ¿Está de acuerdo?

6. (\*) Este ejercicio proviene del libro de Enseñanza Media 4to Matemática, Gonzalo Riera et al, Ediciones Universidad Católica 2002.

Se desea conocer la opinión de las mujeres adultas sobre un programa nacional de guarderías infantiles. La persona contratada para realizar el estudio consigue el listado de las 250 mujeres de un centro femenino, envía 100 cuestionarios y recibe 61 de vuelta. Discute los sesgos que pueden ocurrir en este muestreo.

7.

<sup>8</sup>TIMMS 2003.

<sup>9</sup>TIMMS 1999.

- (a) Sea  $Z \sim \mathcal{N}(0, 1)$ . Calcule  $\mathbb{P}(Z \leq -0,78)$  y  $\mathbb{P}(Z \geq -3,2)$ .
- (b) Sea  $X \sim \mathcal{N}(4, 2)$ . Calcule  $\mathbb{P}(X \leq 3,1)$  y  $\mathbb{P}(X \geq 2,8)$ .
- (c) Sea  $X \sim \mathcal{N}(4, 2)$ . Encuentre  $u$  tal que  $\mathbb{P}(X \geq u) = 0,10$ .
8. Supongamos que la probabilidad de que una pareja tenga un hijo o una hija es igual. Calcule la probabilidad de que una familia con 6 descendientes tenga 2 hijos. Especifique las distribuciones.
9. Suponiendo un porcentaje en la población de 50 %, se selecciona una muestra de 150 alumnos en un colegio de 2000 alumnos. Para obtener el mismo error estándar en un colegio de 4000 alumnos, ¿cuál es el tamaño de muestra que se debe elegir? 156 ó 300.
10. Para tamaños de población  $N$  y de muestra  $n$  dados, ¿cuál de las proporciones  $p$  en la población produce el mayor Error Cuadrático Medio para la proporción en la muestra: 0,50 ó 0,30?
11. (\*) Este ejercicio proviene del libro de Enseñanza Media 4to Matemática, Gonzalo Riera et al, Ediciones Universidad Católica 2002.
- Para estimar la distancia promedio entre el hogar y la oficina para los empleados de un gran empresa, se obtuvo una muestra aleatoria de tamaño 100. La distancia media en la muestra fue de 12 km con una desviación estándar de 6 km.
- (a) Si te piden adivinar la distancia promedio para todos los empleados de la oficina, ¿cuál sería tu adivinanza?
- (b) Encuentra un margen de error aproximado para esta estimación con un nivel de confianza de 95 %.
- (c) ¿En qué intervalo crees tú que se encuentra la distancia media a nivel de toda la empresa?
12. (\*) Este ejercicio proviene del libro de Enseñanza Media 4to Matemática, Gonzalo Riera et al, Ediciones Universidad Católica 2002. Encuentra el tamaño mínimo necesario que debe tener una muestra aleatoria de personas, para estimar el porcentajes de partidarios de un candidato con un error máximo de 2 % y un 90 % de confianza.
13. En la construcción de un intervalo de confianza de una proporción, se considera el intervalo que tiene una probabilidad  $1 - \alpha$  de contener la proporción verdadera  $p$  de una población de tamaño finito  $N$ , ¿cuál de los dos intervalos es más ancho?, el de nivel de confianza  $1 - \alpha = 95\%$  ó  $1 - \alpha = 99\%$ .
14. En un hospital, se lleva un registro del sexo de los recién nacidos, ¿cuál de los sucesos siguientes le parece que tiene más probabilidad de ocurrir?
- (a) Que entre los próximos 10 recién nacidos haya más de 70 % de niñas.
- (b) Que entre los próximos 100 recién nacidos haya más de 70 % de niñas.
- (c) Las dos afirmaciones son igualmente probables.
15. En el conjunto de los 2400 colegios que rindieron la prueba SIMCE en 2do medio, se tiene una media de 254,6 y una desviación estándar de 49. Se selecciona una m.a.s.

de 300 colegios para aplicar algún programa. Se encuentra en la muestra un promedio de 260 en el SIMCE. Para verificar si la muestra es aceptable construya:

- (a) Un intervalo de confianza a 99 % tomando en cuenta que la población es finita.
  - (b) Un intervalo de confianza a 95 % tomando en cuenta que la población es finita.
  - (c) Un intervalo de confianza a 95 % considerando la población grande (infinita).
  - (d) Compare los intervalos y concluya que pasa entre los casos (b) y (c).
16. Una gran empresa quiere determinar la edad media  $\mu$  de sus clientes. A partir de una muestra aleatoria simple de 100 clientes encuentra un edad promedio de 43 años. Se conoce la desviación estándar que vale  $\sigma = 12$ .
- (a) Encuentre el intervalo de confianza a 95 % para  $\mu$ .
  - (b) Si queremos tener un intervalo de ancho igual a 4 años, ¿cuál debería ser el tamaño de la muestra?
17. La legislación impone a los aeropuertos algunas normas sobre los ruidos emitidos por los aviones en el despegue y aterrizaje. Para las zonas habitadas cercanas del aeropuerto de Pudahuel, el límite tolerado propuesto es de 80 decibeles. Si el aeropuerto no cumple esta norma, tendrá que indemnizar a los afectados. Los habitantes de Quilicura aseguran que el ruido sobrepasa 80 decibeles para ciertos tipos de aviones y el aeropuerto asegura que es solamente de 78 decibeles. Los expertos consultados deciden registrar la intensidad del ruido de estos aviones. Toman una muestra de 200 aviones y obtienen una intensidad media de 79,1 decibeles con una desviación estándar de 7 decibeles.
- (a) Tomando  $H_0 = 80$  y  $H_1 = 78$ , concluya quién tiene la razón con un error  $\alpha$  de 5 %. Dé el otro error  $\beta$ .
  - (b) Tomando  $H_0 = 78$  y  $H_1 = 80$ , ¿llega a la misma decisión con un error  $\alpha$  de 5 %? Dé el otro error  $\beta$ .
  - (c) ¿Qué concluye?, ¿hay una contradicción?, ¿cuáles son las hipótesis que favorecen al aeropuerto?
18. Un fabricante ha desarrollado una nueva cortadora de pasto. Afirma que el nuevo motor es mucho más eficiente que los antiguos y que puede funcionar al menos 5 horas seguidas con 4 litros de bencina. Con una muestra aleatoria de 85 máquinas se obtuvo una media de 294 minutos. La desviación estándar es de 24 minutos en la población.
- (a) Escriba las hipótesis nula y alternativa del test.
  - (b) Concluya si la nueva máquina rinde efectivamente 5 horas con un error de tipo I de 5 %.
  - (c) Calcule el p-valor. Interprete.
19. Un productor de fertilizantes vende sus productos en bolsas de 22 kg. El peso de las bolsas tiene un distribución normal de desviación estándar de 0,3 kg. Un consumidor afirma que la compañía comete un fraude, ya que las 15 bolsas que compró tenían un peso promedio de 21,4 kg.
- (a) Escriba las hipótesis nula y alternativa del test.

(b) ¿La queja del consumidor es justificada?

20. En centro de estudios nucleares un investigador pone 23 peces en un medio radioactivo y mide la talla, la radioactividad de las escamas y de los ojos. Quiere estudiar el efecto del peso de los peces sobre la radiactividad de las escamas y de los ojos<sup>10</sup>.

(a) Formule las hipótesis.

(b) Construya la región crítica para  $\alpha = 5\%$  y calcule el p-valor a partir de la tabla adjunta. Concluya.

(c) Repita las dos partes anteriores para la radioactividad de los ojos.

Radio-actividad	Pez grande			Pez pequeño		
	Media	Desviación estándar	Números peces	Media	Desviación estándar	Números peces
Escamas	121,8	25,2	5	177,7	34,3	6
Ojos	9,4	3,14	5	11,7	5,47	6

<sup>10</sup>Datos extraídos de “Introduction à l’analyse des données”, F Cailliez y J.P. Pàges, SMASH, 1976.



## Capítulo 4: Regresión lineal simple



### 4.1 Introducción a los modelos

Generalmente, un estudio estadístico involucra más de una variable. Por ejemplo, una encuesta de opinión sobre la percepción de un producto contiene varias preguntas, cuyas respuestas se relacionan con la opinión que tiene el encuestado sobre el producto, respuestas que serán muy útiles para la elaboración de una campaña publicitaria. En una encuesta política, se interrogan los votantes, no solamente sobre su candidato preferido, sino también sobre su edad, género, profesión, etc. El análisis de tal encuesta permitirá eventualmente determinar el perfil del electorado del candidato, y, por tanto, orientar su campaña electoral. En un estudio escolar, un psicólogo quiere comparar las aptitudes mentales (CI) y el rendimiento escolar de un grupo de estudiantes. Para implementar el proceso de recuperación de cobre en una mina, el químico toma en cuenta el valor de diferentes variables de las cuales depende, tales como el tiempo de flotación, la granulometría, el Ph y la velocidad angular del rotor. Todos estos estudios llevan a medir, describir y modelar las relaciones existentes entre ellas, para sacar conclusiones útiles con el objeto de proponer nuevas campañas publicitarias o políticas, implementar procesos, etc.

¿Quién no ha deseado alguna vez poder predecir el futuro? Algunas personas tienen tantas ansias de saber qué pasará mañana, que estudian los mapas astrales, las cartas de tarot o una bola de cristal. Por otro lado, en los medios de comunicación nos bombardean de predicciones de resultados electorales, económicos, del tiempo, etc.

Todos estos estudios serios, orientados a explicar fenómenos o hacer predicciones, no usan una bola de cristal, sino se basan en general en modelos matemáticos obtenidos sobre la base de informaciones pasadas.

Estos modelos, salvo cuando se basan en leyes de la física o de la mecánica, no son exactos, pues en general, no pueden tomar en cuenta todos los aspectos del problema y simplifican la realidad.

Muchas veces en el estudio de datos de distintas variables de un mismo fenómeno, observamos que es posible encontrar una cierta dependencia entre ellas. En el gráfico de dispersión dado en la solución del ejercicio 10 del Capítulo 2(Figura (a)), que representa el peso y la estatura de algunos niños, observamos que para estos niños, a mayor estatura, mayor peso. Los puntos no se ajusten exactamente a ninguna función simple conocida (lineal, cuadrática, exponencial, etc.), pero vemos que tienden a alinearse. Se habla, en este caso, de **tendencia lineal**. Se podría, entonces, trazar una

recta de la cual los puntos no se alejan mucho. Esta recta constituye un modelo para los datos. Cuando el modelo es “bueno” podrá servir para predecir el valor de una variable a partir de la otra. Cuando la pendiente de la recta es positiva, se dice que la tendencia es **positiva** en el sentido que cuando crece una variable, crece la otra. Podemos, entonces, buscar ajustar los puntos con una recta, de manera que ésta pase lo más cerca de los puntos. Si queremos hacer una predicción del peso de un niño para el cual conocemos su estatura H, podemos usar estos datos. La estatura H podría no estar en los datos de los 12 niños (por ejemplo 110 cm), o bien, podría haber varios niños que tienen la misma estatura (el caso de 112 cm) pero no el mismo peso. Con esta recta podremos dar un valor único a la predicción del peso de un niño que tiene una estatura dada. Por ejemplo, un niño de 127 cm puede pesar más o menos 26 kg.

Se puede poner en evidencia una eventual relación existente entre dos variables, realizando el gráfico de dispersión. Consideraremos ahora 106 países y tres variables: tasa de alfabetización ( $X$ ), tasa de mortalidad infantil ( $Y$ ) y  $Z$ , el producto nacional bruto (PNB). En el gráfico de la Figura 4.1(a) vemos una tendencia lineal negativa; a mayor tasa de alfabetización, menor tasa de mortalidad infantil. La recta en rojo muestra una aproximación lineal posible, que permitirá predecir, “de manera aproximada”, la tasa de alfabetización de un país del cual conocemos la tasa de mortalidad infantil o viceversa. Para llegar a esta recta, basta encontrar los dos coeficientes  $a$  y  $b$  de la recta a partir de los 106 valores de las variables  $X$  e  $Y$  de los 106 países:  $Y = aX + b$ . Por ejemplo, sabemos que un país tiene una tasa de alfabetización de 30 %, pero no conocemos su tasa de mortalidad infantil. Se puede predecir su tasa de mortalidad usando la recta:  $Y = aX + b$ . Será aproximadamente 115 muertes por 1000 nacimientos.

El gráfico de dispersión del PNB con la tasa de mortalidad infantil (Figura 4.1(b)), no muestra una tendencia lineal, pero presenta un aspecto más bien hiperbólico ( $Y = a/Z$ ). Una vez obtenida esta función  $f : \mathbb{R} \rightarrow \mathbb{R}$ , podemos hacer predicciones para nuevos países de la variable  $Y$  o  $Z$  conociendo la otra.

Es nuevamente el modo de razonamiento del estadístico, el que prefiere sacrificar un poco de información (los puntos no están sobre una recta u otra curva en general, pero se desvían eventualmente poco de una curva) para obtener un modelo que es más fácil de usar y facilita la interpretación de los datos. Pero la elección de un tipo de curva tiene que tomar en cuenta que **el modelo tiene que adaptarse a los datos y no al revés**.

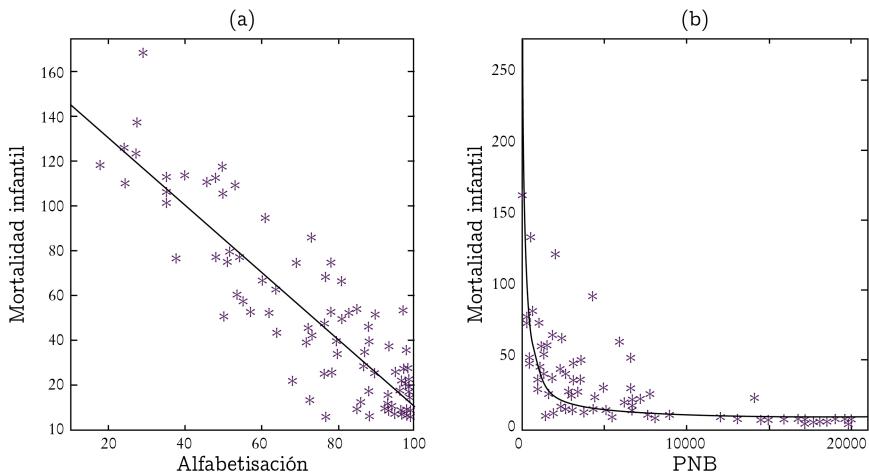
Si conocemos dos variables, por ejemplo  $X$  y  $Z$ , la predicción de la variable  $Y$  podría eventualmente mejorarse utilizando ambas variables  $X$  y  $Z$  en el modelo. Si es una función lineal:  $Y = aX + bZ + c$ .<sup>1</sup>

Previo a la construcción de la recta, conviene saber ¿qué tan fuerte es la tendencia lineal entre las dos variables? Si no existe un cierto grado de relación lineal, no tiene

---

<sup>1</sup>Esta monografía trata solamente el caso de una variable explicativa con un modelo lineal. Se llama regresión lineal simple.

FIGURA 4.1. Ejemplos de los países



sentido buscar una recta que no será de mucha utilidad. La intensidad de la relación lineal puede medirse mediante los conceptos de covarianza y de correlación.

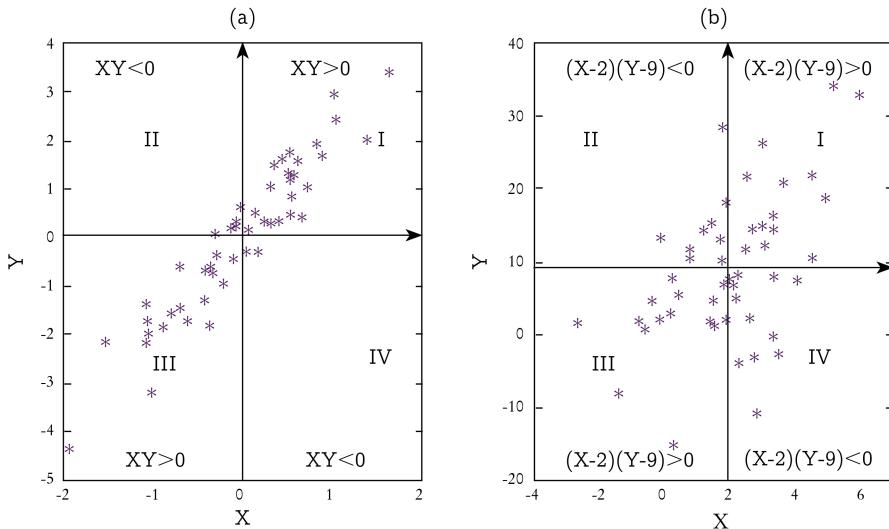
#### 4.2 Coeficiente de correlación lineal

El coeficiente de correlación lineal, debido a Karl Pearson (1857- 1936) permite medir el grado de aproximación del alineamiento de los puntos en el plano.

Consideramos el gráfico de dispersión 4.2(a) de 50 puntos  $(X_i, Y_i)$ ,  $(i = 1, 2, \dots, 50)$ , cuyas medias de  $X$  e  $Y$  son nulas. Dividimos el plano en 4 cuadrantes y vemos que en los cuadrantes I y III el producto  $XY$  es positivo y en los dos otros cuadrantes es negativo. Muy pocos productos son negativos, lo que indica una tendencia positiva. Una recta que ajustará bien los puntos debe, entonces, tener una pendiente positiva.

En el gráfico 4.2(b), la media de  $X$  es 2 y la de  $Y$  es 9. En este caso, se construye los cuadrantes a partir de los ejes intersectándose en el punto  $(2,9)$ . Consideramos ahora los productos “centrados”,  $(X - 2)(Y - 9)$ . Existe una tendencia positiva también, pero se observa más productos negativos que en el caso anterior, lo que hace la relación menos clara.

FIGURA 4.2. Signos de  $(X - M_X)(Y - M_Y)$



Una medida de asociación que expresa el grado de influencia que tienen dos variables entre sí y que se puede expresar mediante la ecuación de una recta, se basa en estos productos. Sea un conjunto de  $n$  pares de datos  $\{(X_i, Y_i) | i = 1, 2, \dots, n\}$ . Denotamos  $M_X$  y  $M_Y$  las medias de los  $X_i$  e  $Y_i$ , respectivamente. Una primera medida de la relación es la covarianza:

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (X_i - M_X)(Y_i - M_Y).$$

Propiedades de la covarianza, que se deja comprobar como ejercicio:

- $Cov(X, X) = \text{Var}(X)$ .
- Si  $U = c_1X + c_2$ , entonces  $Cov(U, Y) = c_1Cov(X, Y)$  y  $Cov(Y, U) = c_1Cov(X, Y)$ .
- Si  $U = c_1X + c_2$  y  $V = d_1Y + d_2$ , entonces  $Cov(U, V) = c_1d_1Cov(X, Y)$ .
- Sean tres variables,  $X$ ,  $Y$  y  $Z$ , entonces  $Cov(X, Y + Z) = Cov(X, Z) + Cov(Y, Z)$ .

La covarianza presenta el inconveniente de ser difícil de interpretar, ya que depende de las unidades de medición. De hecho, tiene su propia unidad de medición, que es el producto de las unidades de las dos variables. Por ejemplo, no sabemos si el valor 3 es grande o pequeño para una covarianza. De hecho, si  $X$  se mide en metros e  $Y$  en kilogramos y si la covarianza entre  $X$  e  $Y$  vale 3, el cambio de la unidad de  $X$  a centímetro, produce una covarianza de 300, sin que cambie la influencia que pueden tener las dos variables entre sí.

Se usa, entonces, un índice que permite eliminar el efecto de las escalas de medición y que puede interpretarse mejor. El coeficiente de correlación lineal se obtiene normalizando la covarianza:

$$r_{X,Y} = \frac{\sum_{i=1}^n (X_i - M_X)(Y_i - M_Y)}{\sqrt{\sum_{i=1}^n (X_i - M_X)^2} \sqrt{\sum_{i=1}^n (Y_i - M_Y)^2}} = \frac{Cov(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}.$$

A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medición de las variables. Tenemos las siguientes propiedades del coeficiente de correlación:

- $-1 \leq r_{X,Y} \leq +1$ .
- $r_{X,Y}$  es simétrico en  $X$  e  $Y$ :  $r_{X,Y} = r_{Y,X}$ .
- $r_{X,Y} = 1$  indica que existe una dependencia lineal perfecta con pendiente positiva.
- $r_{X,Y} = -1$  indica que existe una dependencia lineal perfecta con pendiente negativa.
- $r_{X,Y} > 0$  y cercano a 1 indica una tendencia lineal positiva, es decir, a mayor valor de  $X$ , mayor valor de  $Y$  y viceversa, pero la relación no es lineal.
- $r_{X,Y} < 0$  y cercano a -1 indica una tendencia lineal negativa, es decir, a mayor valor de  $X$ , menor valor de  $Y$  y viceversa, pero la relación no es lineal.
- $r_{X,Y} = 0$  indica que no existe dependencia lineal. Pudiera existir otro tipo de dependencia.

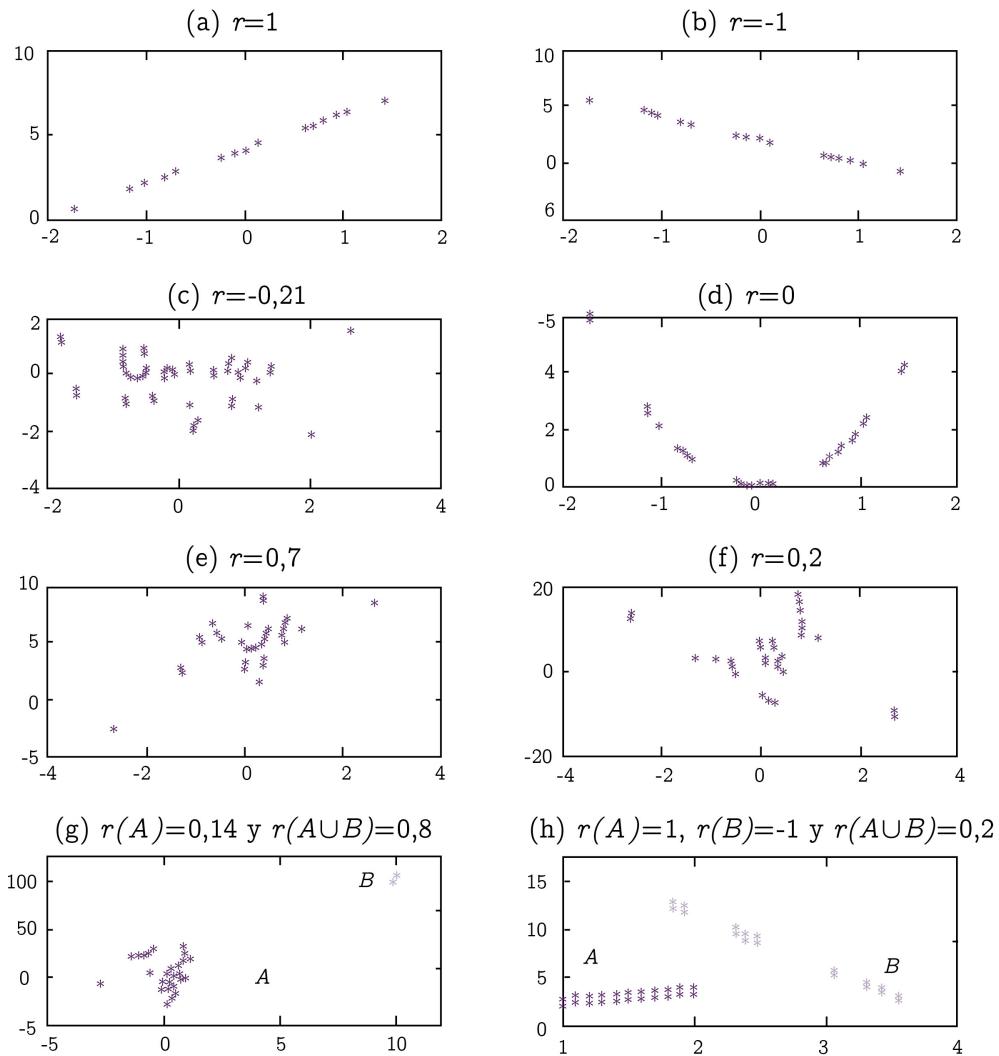
A modo de ejercicio, trate de verificar las 4 primeras propiedades. En resumen, el coeficiente de correlación es un índice que varía entre -1 y 1. Crece de 0 a 1 cuando crece la tendencia lineal positiva y decrece de 0 a -1 cuando crece la tendencia lineal negativa. El coeficiente de correlación toma generalmente valores distintos de -1, 1 ó 0, lo que dificulta su interpretación. Por ello, se debe tener presente que la interpretación de los coeficientes de correlación son dependientes del estudio realizado. Por ejemplo, en algunos experimentos de laboratorio biológico o químico bien controlados, no es difícil encontrar coeficientes de correlación altos, en contraste, en estudios de ciencias humanas, psicológicas o económicas se observan habitualmente coeficientes no muy altos. Por ejemplo, los coeficientes de correlación del rendimiento en la universidad con los resultados en la PSU, si bien son generalmente positivos, son menores que 0,3, lo que indica una tendencia positiva pero no muy fuerte.

¿Un coeficiente de correlación cercano a +1 ó -1 siempre corresponde una tendencia lineal? Examine y comente los gráficos de las Figuras 4.3.

Se desprende un consejo: junto con un coeficiente de correlación realice un gráfico de dispersión de las dos variables para verificar la presencia de puntos atípicos (Figura 4.3(g)) o existencia de una mezcla de poblaciones distintas (Figura 4.3(h)).

Es frecuente encontrarse con una tendencia no lineal (Figura 4.3(d)). En ciertos casos se puede medir el grado de relación no lineal, utilizando el coeficiente de correlación lineal después de transformar de manera adecuada las variables para aproximarse

FIGURA 4.3. Gráficos de dispersión y coeficientes de correlación



a una relación lineal. En el gráfico (d),  $r_{X^2,Y} = 1$ . En el caso del ejercicio 10, hay que tomar  $1/Y$  en vez de  $Y$  en el coeficiente de correlación lineal.

Es muy común inferir de una relación de correlación, una relación de causalidad. Veamos, a través de ejemplos, lo que significa un coeficiente de correlación lineal, que no necesariamente implica que una variable es “causa” de la otra.

- Hay una correlación lineal alta positiva entre el número de horas de sueño y el tiempo que vive una persona, por lo tanto, las personas que duermen solamente 5 ó 6 horas por noche, viven más tiempo que las personas que duermen más. ¿Esto significa que tiene que dormir menos para vivir más?, ¿no será que un tercer elemento, el modo de vida de la persona, explica esta correlación? En efecto, las personas activas duermen menos y viven más tiempo. Podría ser que el modo de vida de la persona es causa del tiempo de sueño y del tiempo de vida.
- ¿Han aumentado los ingresos de los habitantes de Zedlandia en las últimas décadas o han disminuido? <sup>2</sup> La media de ingresos monetarios por hogar ha descendido: en 1970 ascendía a 34.200 zeds, en 1980 era de 30.500 zeds y en 1990 de 31.200 zeds. No obstante, los ingresos por persona aumentaron: en 1970 ascendieron a 13.500 zeds, en 1980 fueron de 13.850 zeds y en 1990 de 15.777 zeds. Un hogar está formado por todas las personas que viven juntas en una misma vivienda. Explica cómo es posible que en Zedlandia desciendan los ingresos por hogar a la vez que aumentan los ingresos por persona.
- El polígrafo es conocido como detector de mentiras. Tiene una correlación del orden del 88% en la detección de mentiras. En realidad, lo que detecta el polígrafo son alteraciones fisiológicas generadas por la activación emocional del individuo cuando hay una divergencia entre lo que dice y lo que siente. Hay una gran correlación entre detectar mentiras y la casualidad de detectar alteraciones en el cuerpo humano; pero, aunque el porcentaje de aciertos es muy elevado, carece de rigor científico.
- Se encontró una correlación positiva entre el consumo de helado y los ahogos en el mar. No será que en verano uno se baña en el mar y se consume más helados que en invierno.
- La mayoría de los accidentes automovilísticos ocurren con vehículos en velocidad moderada y hay muy pocos accidentes con vehículos que transitan a alta velocidad, ¿esto indicaría un coeficiente de correlación lineal entre la velocidad del vehículo y el número de accidentes negativo?, ¿esto significa que es más seguro andar a alta velocidad? Aquí se debe considerar la tasa de accidente para cada nivel de velocidad. Se encontraría lo que uno espera: a mayor velocidad, mayor probabilidad de tener un accidente. En este caso es bastante razonable hablar de causalidad: la alta velocidad es posiblemente una causa de accidente. Pero, seguramente hay otras causas.

En resumen, dos fenómenos correlacionados, no implica que uno es la causa del otro. Las causas requieren más información que un coeficiente de correlación. Se busca con un trabajo científico más profundo.

<sup>2</sup>Pregunta de la prueba de matemática PISA 2006

### 4.3 Planteamiento de la regresión lineal simple

Anteriormente vimos cómo medir el grado de relación de tipo lineal entre dos variables a partir del coeficiente de correlación lineal y que éste es un índice simétrico. Sin embargo, rara vez los roles de las variables son simétricos. El ph puede influir sobre la recuperación del mineral, pero la recuperación del mineral no influye sobre el ph. Una variable  $X$  puede influir sobre la variable  $Y$ , pero la recíproca no es necesariamente cierta. Vimos incluso que más de una variable puede influir al mismo tiempo sobre la variable  $Y$ . Quisiéramos, entonces, no solamente evaluar la intensidad de la asociación, sino encontrar también la ecuación del modelo.

Algunas relaciones son conocidas y deterministas como ciertas leyes de la física o de la mecánica, pero dependen de constantes desconocidas que hay que determinar. Estas constantes pueden obtenerse a partir de experimentos que se utilizarán en el modelo ya planteado. El problema que surge, entonces, en la determinación de las constantes, está en los errores de mediciones. Lo que lleva a tomar muchas más mediciones que el número de constantes a estimar.

En otros problemas las relaciones no son conocidas y hay que determinar completamente el modelo. En ciencias sociales o en economía, por ejemplo, los modelos no son deterministas y contienen una componente aleatoria, lo que dificulta la búsqueda de las relaciones. En este caso se quiere descubrir cómo un conjunto de variables influye sobre otra variable. Según el contexto, las variables se llaman de diferentes maneras. Consideramos solamente el modelo con dos variables, con el cual se quiere determinar los valores de la variable  $Y$ , a partir de los valores de la variable  $X$ :

$$Y = aX + b \quad (4.1)$$

Este modelo se llama “modelo de regresión simple”. La idea central de este modelo es que la respuesta media de la variable  $Y$  cambia con los valores de  $X$  y esto de manera proporcional al valor de  $X$ .

Según el contexto de los datos, se designa las variables de diferentes maneras. La variable  $Y$  se llama **variable a explicar, variable respuesta, variable endógena o variable dependiente** y  $X$  se llama **variable explicativa, variable exógena o variable independiente**. Se dice que el modelo es de regresión “simple” cuando tiene una sola variable explicativa.

Los coeficientes  $a$  y  $b$  del modelo son desconocidos y se obtienen (estiman) a partir de los datos empíricos. Por una razón histórica, este modelo se llama *regresión lineal*. Los mayores descubrimientos de Sir Francis Galton fueron sus formulaciones sobre la regresión. En particular, realizó un estudio que mostró que la estatura de los hijos nacidos de padres altos tiende a retroceder o “regresar” hacia la estatura promedio de la población, a pesar de mostrar una tendencia lineal para las alturas medianas. Por lo que utilizó, entonces, la palabra “regresión lineal” para referirse a un modelo del tipo  $Y = aX + b$ , donde  $Y$  es una variable a explicar y  $X$  una variable explicativa.

#### 4.4 Criterio de mínimos cuadrados

Consideramos la estatura  $X$  y el peso  $Y$  con los datos del ejercicio 10. Si la ecuación 4.1 se cumple deberíamos tener dos escalares  $a$  y  $b$  tal que:

$$\begin{cases} 102,5 = a13,6 + b \\ 112,0 = a16,35 + b \\ \vdots \\ 130,2 = a29,55 + b \end{cases}$$

Es un sistema de 12 ecuaciones lineales con dos incógnitas (los parámetros  $a$  y  $b$  del modelo). Como los 12 puntos no son colineales, no tiene solución. Agregamos, entonces, términos de “errores” para encontrar una solución aproximada al sistema de ecuaciones(Figura 4.4).

$$\begin{cases} 102,5 = a13,6 + b + e_1 \\ 112,0 = a16,35 + b + e_2 \\ \vdots \\ 130,2 = a29,55 + b + e_{12} \end{cases}$$

Parece que estamos frente a un problema, ya que, a pesar de seguir teniendo un sistema de 12 ecuaciones lineales, aumentamos el número de incógnitas a 14! Sigue sin solución.

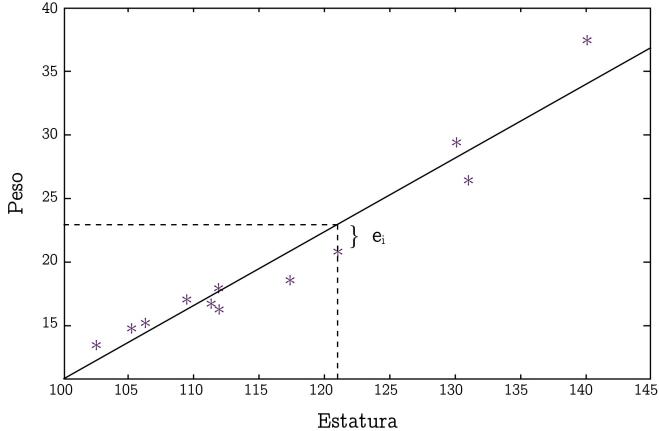
Aplicaremos nuevamente el método de mínimos cuadrados (Sección 3.4.1) que Gauss usó en el análisis numérico de los errores de mediciones en física y astronomía. El método no se publicó hasta 1809, apareciendo en el segundo volumen de su trabajo sobre mecánica celeste, “Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium”. El francés Adrien-Marie Legendre desarrolló el mismo método de forma independiente en 1805.

Pero, los términos de errores que agregamos no tienen el mismo rol que los parámetros  $a$  y  $b$  del modelo. Queremos que los errores sean lo más pequeño posible. Naturalmente no podemos hacerlos todos nulos. Pero podemos minimizar una función de los errores  $h(e_1, e_2, \dots, e_{12})$ . La función no puede ser la media de los errores. ¿Por qué? Pues los errores pueden ser positivos o negativos pero grandes en valor absoluto. Entonces, podemos minimizar el promedio de los valores absolutos de los errores  $h = \frac{1}{12} \sum_{i=1}^{12} |e_i|$ . La solución a este problema de optimización, que consiste en encontrar  $a$  y  $b$  que minimizan la función  $h$ , no es muy simple numéricamente. Si bien es muy interesante, se usa menos en la práctica. La función más usual que se minimiza es cuadrática:  $h = \frac{1}{12} \sum_{i=1}^{12} e_i^2$ . Se llama **Criterio de los Mínimos Cuadrados**.<sup>3</sup>

---

<sup>3</sup>Vimos en el tercer capítulo 3.4.1 una nota histórica del método de los Mínimos Cuadrados.

FIGURA 4.4. Errores del modelo



La solución es explícita. En efecto, para minimizar  $h$  con respecto de  $a$  y  $b$  basta anular las dos derivadas parciales de  $h = \sum_{i=1}^{12} (y_i - aX_i - b)^2$ :

$$\begin{cases} \frac{\partial h}{\partial a} = -2 \sum_{i=1}^{12} (Y_i - aX_i - b)X_i = 0 \\ \frac{\partial h}{\partial b} = -2 \sum_{i=1}^{12} (Y_i - aX_i - b) = 0 \end{cases}$$

De la segunda ecuación obtenemos:  $b = M_Y - aM_X$ , donde  $M_X$  y  $M_Y$  son las medias de los  $X_i$  e  $Y_i$ , respectivamente. Se reemplaza  $b$  en la primera ecuación:

$$a \sum_{i=1}^{12} X_i^2 = \sum_{i=1}^{12} X_i Y_i - (M_Y - aM_X) \sum_{i=1}^{12} X_i$$

$$a \left( \sum_{i=1}^{12} X_i^2 - M_X \sum_{i=1}^{12} X_i \right) = \sum_{i=1}^{12} X_i Y_i - M_Y \sum_{i=1}^{12} X_i$$

Usando  $\sum_{i=1}^n X_i = nM_X$  y  $\sum_{i=1}^n Y_i = nM_Y$ , obtenemos:

$$a \left( \sum_{i=1}^{12} X_i^2 - nM_X^2 \right) = \sum_{i=1}^{12} X_i Y_i - nM_Y M_X$$

Notemos que  $\text{Var}(X)$ ,  $\text{Var}(Y)$  y  $\text{Cov}(X, Y)$  pueden escribirse como:

$$\left\{ \begin{array}{l} \text{Var}(X) = \frac{1}{n} \sum_{i=1}^n X_i^2 - M_X^2 \\ \text{Var}(Y) = \frac{1}{n} \sum_{i=1}^n Y_i^2 - M_Y^2 \\ \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n X_i Y_i - M_X M_Y \end{array} \right.$$

Denotamos  $\hat{a}$  y  $\hat{b}$  la solución de los mínimos cuadrados para  $a$  y  $b$ , respectivamente. Obtenemos, entonces:

$$\left\{ \begin{array}{l} \hat{a} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \\ \hat{b} = M_Y - \hat{a} M_X \end{array} \right.$$

Tenemos, entonces, el modelo “estimado” por el criterio de los mínimos cuadrados, que define la recta de regresión:

$$\hat{Y}_i = \hat{a} X_i + \hat{b}, \quad i = 1, 2, \dots, n \quad (4.2)$$

Los errores son, entonces, estimados con las diferencias entre los valores observados y los valores estimados de  $Y$ :  $\hat{e}_i = Y_i - \hat{Y}_i$ , errores denominados “residuos”, para distinguirlos de los errores desconocidos del modelo antes de estimar los coeficientes  $a$  y  $b$ . Estos se llaman “errores teóricos”. El modelo después de estimar se escribe:

$$Y_i = \hat{Y}_i + \hat{e}_i = \hat{a} X_i + \hat{b} + \hat{e}_i, \quad i = 1, 2, \dots, n \quad (4.3)$$

Dos resultados importantes:

- La media de los residuos es nula:  $\sum_{i=1}^n \hat{e}_i = 0$ . Se debe a que la media de los  $\hat{Y}_i$  es igual a la media de los  $Y_i$ . En efecto

$$M_{\hat{Y}} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i = \frac{\hat{a}}{n} \sum_{i=1}^n X_i + \hat{b} = \hat{a} M_X + \hat{b} = M_Y$$

- Mostramos que  $\sum_{i=1}^n \hat{Y}_i \hat{e}_i = 0$ , o sea el vector formado de los valores de  $\hat{Y}_i$  y el vector formado de los residuos  $\hat{e}_i$  son vectores ortogonales de  $\mathbb{R}^n$ :

$$\sum_{i=1}^n \hat{Y}_i \hat{e}_i = \sum_{i=1}^n (Y_i - \hat{Y}_i) \hat{Y}_i = \sum_{i=1}^n Y_i \hat{Y}_i - \sum_{i=1}^n \hat{Y}_i^2 = \hat{a} \sum_{i=1}^n Y_i X_i + \hat{b} \sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i^2$$

Sabemos que  $\sum_{i=1}^n \hat{Y}_i = \sum_{i=1}^n Y_i$ , luego,

$$\sum_{i=1}^n \hat{e}_i \hat{Y}_i = \hat{a} \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n Y_i (\hat{b} - \hat{Y}_i) = \hat{a} \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n \hat{Y}_i (\hat{b} - \hat{Y}_i)$$

Como  $\hat{b} - \hat{Y}_i = \hat{a} X_i$ ,

$$\sum_{i=1}^n \hat{e}_i \hat{Y}_i = 0. \quad (4.4)$$

Aplicamos estos resultados al ejemplo de los 106 países, tomando la alfabetización como variable explicativa y la mortalidad infantil como variable a explicar. Las medias y varianzas de las dos variables se encuentran en la Tabla 4.1 y el coeficiente de correlación lineal es igual a  $-0,892$ . Calculamos los coeficientes después de calcular  $Cov(X, Y) = r_{X,Y} \sqrt{\text{Var}(X) \text{Var}(Y)} = -784,6$ :

$$\begin{cases} \hat{a} = \frac{-784,6}{22,85^2} = -1,503 \\ \hat{b} = M_Y - \hat{a} M_X = 160,6 \end{cases}$$

Se encuentra un coeficiente  $\hat{a}$  negativo, lo que había de esperar, dada la correlación negativa entre las dos variables. Esto nos dice que un país con una tasa de alfabetización alta, tendrá posiblemente una tasa de mortalidad baja y recíprocamente.

TABLA 4.1. Datos de los países

	Alfabetización	Mortalidad infantil
Media	78,22	43,03
Desviación estándar	22,85	38,12

## 4.5 Predicciones

Supongamos que el país “Atlántida” no fue considerado en el ejemplo anterior, debido a que sólo se conocía su tasa de alfabetización de 64 %. Con el modelo encontrado podemos “estimar” o “predecir” la tasa de mortalidad infantil de Atlántida a partir de su tasa de alfabetización, suponiendo que la relación entre los valores es parecida a la de los países que se usaron para estimar  $a$  y  $b$ . Vale decir, que el punto correspondiente a Atlántida estaría cerca de la recta de regresión. Lo natural en este caso es tomar el punto sobre la recta correspondiente a la tasa de alfabetización de 64 %. La predicción de tasa de mortalidad de Atlántida es, entonces:  $-1,503 * 64 + 160,6 = 64,4$ .

Se habla de predicción cuando se obtiene un valor de la variable  $Y$  a partir del modelo estimado 4.2. Lo que no significa que es una predicción en el tiempo.

Para saber si se puede confiar en la predicción de la tasa de mortalidad infantil de Atlántida, tenemos que verificar la validez del modelo que se usó.

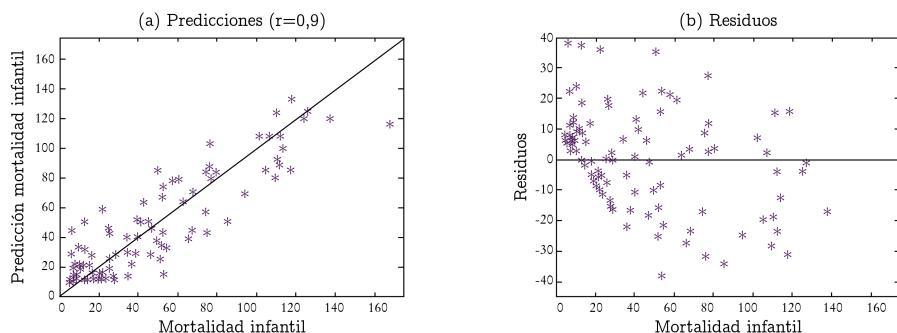
La predicción obtenida del modelo de regresión lineal tiene una propiedad interesante: no depende de las unidades de medición. En efecto, si la recta encontrada para  $X$  e  $Y$  es  $\hat{Y} = \hat{a}X + \hat{b}$ , y si se cambia la variable  $X$  a  $X' = uX$ , entonces, la recta para  $Y$  y  $X'$  será:  $\hat{Y}' = \hat{a}'X' + \hat{b}' = \hat{a}\frac{X'}{u} + \hat{b}$ . El valor del término  $\hat{a}X$  no cambia:  $\hat{a}X = \hat{a}'X'$  y el término constante no cambia; el coeficiente  $\hat{a}'$  es igual a  $\frac{\hat{a}}{u}$ . Si además, se cambia  $Y$  por  $Y' = vY$ , la recta de regresión para  $Y'$  y  $X'$  es:  $\hat{Y}' = \hat{a}''X' + \hat{b}''$ .  $\hat{Y}'$  tendrá el mismo cambio de unidad que  $Y$ , entonces:  $\hat{Y}' = \hat{a}\frac{X'}{u} + \hat{b} = \hat{a}\frac{v}{u}X + v\hat{b}$ . Con las variables  $Y'$  y  $X'$ , la pendiente de la recta es  $\hat{a}'' = \hat{a}\frac{v}{u}$  y  $\hat{b}'' = v\hat{b}$ , pero la recta no cambia.

#### 4.6 Validación del modelo

Para comprobar la validez del modelo, los software estadísticos proporcionan resultados de tests de hipótesis. Estos tests se apoyan sobre estadísticos (funciones de valores muestrales) que no son parte de esta monografía<sup>4</sup>. Sin embargo, es posible usar gráficos de dispersión y el coeficiente de correlación para ver la calidad del modelo.

En efecto, si el modelo es “bueno”, significa que reproduce bien los valores de  $Y$  para todos los valores de  $X$ , o sea  $\hat{Y}$  e  $Y$  discrepan poco. El gráfico de dispersión de  $(Y, \hat{Y})$  debería mostrar los puntos no muy alejados del primer cuadrante. Pero

FIGURA 4.5. Gráficos para validez del modelo



sabemos cuánto vale este coeficiente de correlación. En efecto, tenemos los siguientes resultados:

- $Cov(\hat{e}, \hat{Y}) = \frac{1}{n} \sum_{i=1}^n (\hat{e}_i - 0)(\hat{Y}_i - M_Y) = 0$ .

---

<sup>4</sup>Se presentarán estos tests en otra monografía.

- El coeficiente de correlación de  $Y$  con  $\hat{Y}$  es igual a  $\frac{\sqrt{\text{Var}(\hat{Y})}}{\sqrt{\text{Var}(Y)}}$ . En efecto, de las propiedades de la covarianza y de la ecuación 4.4, obtenemos

$$\text{Cov}(Y, \hat{Y}) = \text{Cov}(\hat{Y} + \hat{e}, \hat{Y}) = \text{Cov}(\hat{Y}, \hat{Y}) + \text{Cov}(\hat{e}, \hat{Y}) = \text{Var}(\hat{Y}).$$

Luego,

$$r_{Y,\hat{Y}} = \frac{\text{Cov}(Y, \hat{Y})}{\sqrt{\text{Var}(Y)} \sqrt{\text{Var}(\hat{Y})}} = \frac{\text{Var}(\hat{Y})}{\sqrt{\text{Var}(Y)} \sqrt{\text{Var}(\hat{Y})}} = \frac{\sqrt{\text{Var}(\hat{Y})}}{\sqrt{\text{Var}(Y)}}.$$

- El coeficiente de correlación lineal entre  $Y$  y  $\hat{Y}$ ,  $r_{Y,\hat{Y}}$ , se llama “coeficiente de correlación múltiple”. Se tiene  $r_{Y,\hat{Y}} = r_{X,Y}$ .

$$r_{Y,\hat{Y}} = \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - M_Y)(\hat{Y}_i - M_{\hat{Y}})}{\sqrt{\text{Var}(Y)} \sqrt{\text{Var}(\hat{Y})}}$$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (Y_i - M_Y)(\hat{Y}_i - M_{\hat{Y}}) &= \frac{1}{n} \sum_{i=1}^n (Y_i - M_Y)(\hat{a}X_i + \hat{b} - M_Y) = \frac{1}{n} \sum_{i=1}^n (Y_i - M_Y)(\hat{a}X_i + \hat{b} - \hat{a}M_X - \hat{b}) \\ &= \frac{\hat{a}}{n} \sum_{i=1}^n (Y_i - M_Y)(X_i - M_X) = \hat{a} \text{Cov}(X, Y). \end{aligned}$$

De la ecuación 2.1 se deduce:

$$\text{Var}(\hat{Y}) = \text{Var}(\hat{a}X + \hat{b}) = \hat{a}^2 \text{Var}(X)$$

Luego,

$$r_{Y,\hat{Y}} = \frac{\hat{a} \text{Cov}(X, Y)}{\hat{a} \sqrt{\text{Var}(Y)} \sqrt{\text{Var}(X)}} = r_{X,Y}$$

Nota: Este último resultado es cierto solamente para la regresión lineal simple.

#### 4.7 Regresión de $Y$ sobre $X$ y regresión de $X$ sobre $Y$

Los roles de las variables en el modelo de la regresión no son simétricos. Cuando se lleva a cabo la predicción se supone que la variable explicativa es conocida, mientras que la variable a explicar no lo es. Sean los datos de la Tabla 4.2 cuyo gráfico de dispersión se encuentra en la Figura 4.6. Si tomamos  $X$  como variable explicativa, el criterio de los mínimos cuadrados se aplica a los errores tomados paralelamente al eje de la variable a explicar  $Y$ . Esta regresión se llama **regresión lineal de  $Y$  sobre  $X$** .

Si ahora queremos predecir la variable  $X$  a partir de la variable  $Y$ , tendremos que aplicar el criterio de los mínimos cuadrados a errores tomados paralelamente al eje de la variable  $X$  (Figura 4.6(b)). Es la **regresión de  $X$  sobre  $Y$** .

Obviamente, el coeficiente de correlación múltiple es el mismo para las dos rectas, ya que es igual al coeficiente de correlación lineal entre  $X$  e  $Y$ , que vale 0,68.

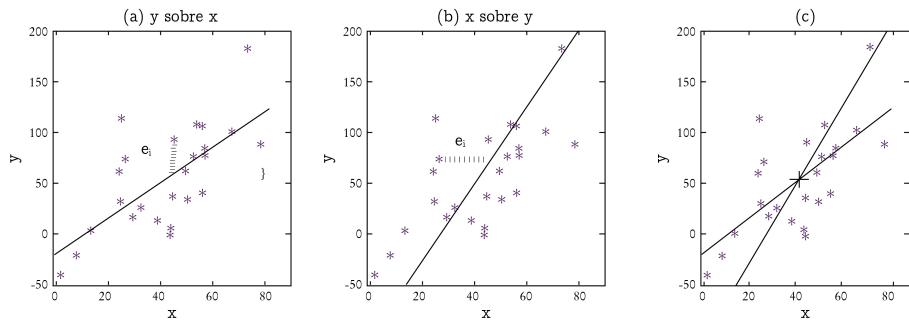
Verifique que para la regresión de  $Y$  sobre  $X$ :  $Y = 1,74X - 18,42$  y para la regresión de  $X$  sobre  $Y$ :  $X = 0,26Y + 27,47$ , o sea  $Y = \frac{X}{0,26} - \frac{27,47}{0,26} = 3,84X - 106,42$ .

Las rectas son bastante diferentes en este caso. Observe que intersectan en el punto de coordenadas de las dos medias ( $M_X, M_Y$ ). Calcule el valor del ángulo formado por las dos rectas. Con un coeficiente de correlación lineal más alto, las dos rectas formarían un ángulo más pequeño. Las dos rectas están confundidas solamente en el caso de un coeficiente de correlación lineal igual a +1 ó -1.

TABLA 4.2. Ejemplo

$X$	12,92	50,15	23,87	73,30	28,90	55,57	49,39	26,56	1,59	43,82	24,79	57,29	55,15
$Y$	2,84	33,04	60,69	184,38	16,02	41,02	61,76	72,99	-40,33	5,66	31,19	82,67	107,00
$X$	78,85	56,83	32,13	52,61	24,82	44,61	44,03	45,00	38,64	66,90	7,52	53,56	
$Y$	88,50	77,76	26,45	74,94	113,98	36,59	-1,31	90,97	12,79	102,14	-20,10	107,11	

FIGURA 4.6. Las dos rectas de regresión



## 4.8 Resumen de la terminología

Coeficiente de correlación lineal: Índice estadístico que mide la relación lineal entre dos variables cuantitativas. No indica causalidad.

Regresión lineal simple: Es un método matemático que modeliza de manera lineal la relación de una variable numérica a partir de otra. No es un modelo simétrico entre las dos variables.

Variable a explicar o dependiente: Es una variable que depende del valor que toman otras variables.

Variable explicativa o independiente: Los cambios en los valores de una variable

explicativa determina cambios en los valores de otra (variable dependiente).

Errores del modelo: Es la diferencia entre los valores observados de la variable a explicar con los valores estimados por el modelo.

Residuos: Son las estimaciones de los errores del modelo.

Criterio de Mínimos cuadrados: Es la función de los errores teóricos del modelo que se minimiza para estimar los parámetros del modelo.

Predictión: Estimaciones de la variable a explicar que se obtienen de un modelo.

## 4.9 Ejercicios

Los ejercicios con \* pueden utilizarse con los estudiantes de Enseñanza Media.

1. (\*) La tabla adjunta recoge el número de horas semanales (variable  $X$ ) que 12 personas dedican a hacer deporte y el número de pulsaciones por minuto (variable  $Y$ ) que tienen en reposo.

Nº horas deporte	0	0	0	1	1	3	3	3	4	5	7	8
Pulsaciones	66	62	73	72	65	60	62	66	58	57	54	55

- (a) Realice un gráfico de dispersión de los datos.  
(b) ¿Hay dependencia funcional entre el número de horas de deporte y el número de pulsaciones?  
(c) La ecuación de la recta de regresión es:  $y = -1,85x + 68,1$ . Estime el número de pulsaciones que tendrá una persona que dedica 2 horas semanales a hacer deporte.  
(d) Estime el número de pulsaciones de un atleta que entrena 4 horas diarias para la maratón de Santiago. ¿Le parece razonable esta estimación? Justifique su respuesta.
2. La pendiente de la recta de regresión de  $Y$  sobre  $X$ , dada por el criterio de los mínimos cuadrados es:  $\hat{a} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$ .
  - (a) Verifique que puede escribirse:  $\hat{a} = \sqrt{\frac{\text{Var}(X)}{\text{Var}(Y)}} r_{x,y}$ .
  - (b) Interprete el coeficiente  $\hat{a}$  a partir de esta nueva formula.

3. (\*) El encargado de finanzas de una empresa estudia la rentabilidad de sus gastos en publicidad. Para ello ha recogido datos del volumen de ventas y del gasto en publicidad expresados en millones de pesos durante 10 años.

- Especifique y estime el modelo lineal que explique las ventas de la empresa en función del gasto en publicidad. Interprete.
- Verifique que la suma de los residuos es nula.
- En el año 2006, la empresa va a invertir 550.000.000 pesos en publicidad. Calcule el volumen de ventas esperado.
- Si se elimina ahora el parámetro constante del modelo de regresión. Estime la pendiente de la recta de regresión y compruebe que la media de los residuos no es nula.

Año	Ventas	Gasto publicidad
1995	50	10
1996	100	15
1997	150	18
1998	200	20
1999	200	25
2000	300	35
2001	400	50
2002	500	55
2003	650	60
2004	700	65

4.

- ¿Por qué muchas veces se usa la recta solución de los mínimos cuadrados para hablar de la regresión lineal? ¿Es correcto?
- Describa los pasos del cálculo de la recta de regresión de  $Y$  sobre  $X$ .
- Los datos del gráfico 4.7(a) entregan un coeficiente de correlación múltiple igual a 0,82. ¿Creen que será útil para hacer la predicción de  $Y$  con  $X = 50$ ?

5. Un economista investiga si existe alguna relación entre el ingreso per cápita (cientos de US dólares) y el porcentaje de la fuerza laboral agrícola a partir de los datos de 15 países en desarrollo (Tabla adjunta).

- Calcule los coeficientes de la regresión lineal del ingreso per cápita sobre el porcentaje de la fuerza laboral agrícola. Comente la ecuación de la recta.
- Si la suma de los cuadrados de los residuos es igual a  $\hat{e}_i^2 = 719,47$ , deduzca la varianza de los residuos.
- Valide el modelo y concluya.
- Si se expresa el ingreso per cápita sin dividirlo por 100 (los valores de la primera columna de la tabla multiplicados por 100). Entregue la recta de regresión con esta nueva unidad para medir la variable  $Y$ , sin usar los datos de la tabla adjunta.

Ingresa per cápita	6	8	8	7	7	12	9	8	9	10	10	11	9	10	11
Fuerza laboral	9	10	8	7	10	4	5	5	6	8	7	4	9	5	8

6. Un médico trata de informar a sus pacientes operados del hígado del tiempo que podrán sobrevivir. Estas predicciones la quiere hacer a partir de los resultados de un examen del hígado realizado a pacientes ya operados. De 60 pacientes antiguos, obtiene informaciones del examen del hígado y del tiempo de sobrevivencia en meses (Tabla adjunta).

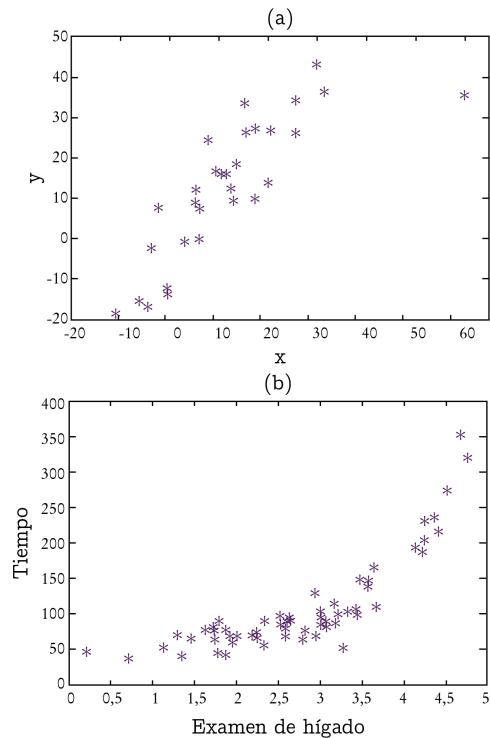
Hígado	3,04	4,21	2,94	2,16	4,49	3,63	2,99	3,46	2,20	3,66	3,03	3,41	1,62	4,39	3,21	0,69	3,25	4,11	3,43	4,22
Tiempo	84	208	61	63	282	165	88	146	56	108	74	99	69	217	95	30	44	193	94	231
Hígado	4,16	1,77	0,28	4,66	3,16	2,77	2,31	1,10	2,98	1,27	1,85	2,52	3,55	3,15	1,30	4,33	2,78	4,75	1,45	1,91
Tiempo	191	81	91	360	81	55	85	46	78	64	68	78	143	111	31	239	66	328	58	61

Hígado	1,92	2,63	1,70	1,71	0,20	1,85	3,02	2,55	2,56	2,62	2,92	2,21	1,99	3,29	2,29	2,99	1,72	2,59	3,54	1,79
Tiempo	49	84	71	69	41	33	75	60	73	87	126	64	59	97	48	95	55	83	138	34

- (a) Considerando el diagrama de dispersión (Figura 4.7(b)) del examen del hígado con el tiempo de sobrevivencia del grupo de los 60 pacientes del médico y el coeficiente de correlación lineal de 0,82, ¿le recomendaría al médico calcular sus predicciones, a partir de la regresión lineal del tiempo sobre el examen del hígado?
- (b) Calcule la recta de regresión y vea gráficamente si los resultados son satisfactorios.
- (c) El médico propone considerar el logaritmo del tiempo en vez del tiempo en la ecuación de regresión. Justifique por qué.
- (d) Tomar  $Y = \log(Tiempo)$ , efectúe la regresión lineal de  $Y$  sobre el examen del hígado.
- (e) Valide gráficamente esta última regresión.
- (f) Si el examen del hígado de un recién operado vale 3,5, ¿cuál es el pronóstico del paciente?
- 7.
- (a) Estudie la existencia y unicidad de la solución de los mínimos cuadrados de la regresión simple.
- (b) ¿Por qué el diagrama de dispersión de los valores de las dos variables del modelo de regresión lineal simple es importante?

FIGURA 4.7. Recta de regresión y residuos





## Anexo 1: Solución de los ejercicios



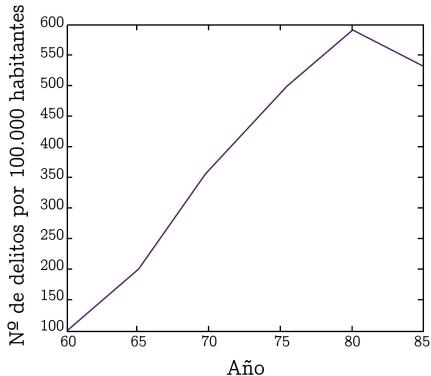
### Capítulo 2

1. Se clasifica una variable  $X$  considerando los valores o alternativas que puede tomar. Si la población es el conjunto  $\mathcal{E}$  y  $\mathcal{Q}$  es el conjunto de los valores posibles que puede tomar  $X$ :  $\mathcal{E} \rightarrow \mathcal{Q}$ .

1. La población es el conjunto  $\mathcal{E}$  de todos los hogares chilenos y  $\mathcal{Q} = \mathbb{R}$ , donde  $\mathbb{R}$  es el conjunto de los números reales. La variable es continua o cuantitativa.
  2. La población  $\mathcal{E}$  es el conjunto de diputados del parlamento chileno del 2008 y  $\mathcal{Q}$  es el conjunto de los partidos políticos representados en el parlamento. Este no es un conjunto numérico, son *nombres*. Luego, la variable es nominal.
  3.  $\mathcal{E}$  es el conjunto de todas las comunas chilenas y  $\mathcal{Q}$  es el número de habitantes de la comuna. Así, la variable es cuantitativa.
  4.  $\mathcal{E}$  es el conjunto de todos los camiones. La carga máxima es una variable cuantitativa y el número de ruedas es una variable discreta.
  5.  $\mathcal{E}$  son todos los hogares chilenos y el consumo anual de electricidad es una variable cuantitativa.
  6.  $\mathcal{E}$  es el conjunto de todos los colegios que rindieron la prueba SIMCE 2º medio en Chile. El promedio SIMCE es una variable cuantitativa.
  7.  $\mathcal{E}$  es el conjunto de todas las galaxias y el número de estrellas es una variable discreta, pero que puede considerarse eventualmente como cuantitativa.
  8.  $\mathcal{E}$  es el conjunto de los niños chilenos menores de 5 años. El género y el color del pelo son variables nominales. El peso y la talla son variables cuantitativas. La edad es una variable cuantitativa que frecuentemente se considera como discreta.
  9.  $\mathcal{E}$  es el conjunto de todos estudiantes de la Universidad de Chile y la carrera es una variable nominal.
- 2.
- (a) Se registraron 100 delitos por cada 100.000 habitantes el año 1960.
  - (b) Los fabricantes de sistemas de alarmas se saltaron los valores de los años 1980, 1981, 1982, 1983 y 1984. Gracias a esto se aprecia un crecimiento de la criminalidad, lo que le permite publicitar a los fabricantes de sistemas de alarma, la venta de su producto.

- (c) El gráfico correcto (considerando los años saltados), que debería emitir la policía, muestra un descenso de la criminalidad (Figura 2.8 ).

FIGURA 2.8. Curva correcta



Los delitos decrecen.

La policía tuvo buen resultado en su lucha contra la delincuencia.

“Hay tres tipos de mentira: las piadosas, las crueles y las estadísticas.” Atribuido a Mark Twain por el primer ministro inglés Benjamin Disraeli (1804-1881).

3. Los resultados son:
  - (a) Inferencia estadística.
  - (b) Inferencia estadística.
  - (c) Estadística descriptiva.
  - (d) Inferencia estadística.
  - (e) Estadística descriptiva.
4. Los resultados son:
  - La mediana no cambia.
  - Los cuartiles no cambian.
5. Los resultados son:
  - Recorrido:  $14 - 5 = 9$ ; Media:  $9,85$ ; Mediana:  $10$ ; Desviación estándar:  $2,17$ .
  - El recorrido no cambia; la media cambia:  $9,85 + 3 = 12,85$ ; la mediana cambia:  $10 + 3 = 13$ ; la desviación estándar no cambia.
  - Todas las medidas son multiplicadas por 2. El recorrido:  $2 \times 9 = 18$ ; la media:  $2 \times 9,85 = 19,7$ ; la mediana:  $2 \times 10 = 20$ ; la desviación estándar:  $2 \times 2,17 = 4,34$ .
6. Los resultados son:
  - (a) La 6º prueba tiene el promedio más alto, con valor 5,7.
  - (b) La 6º prueba tiene las notas más concentradas, puesto que posee una menor desviación estándar, con valor 0,75.
  - (c) Hay que mirar la matriz de notas columna por columna. La 1º prueba es la más simétrica.

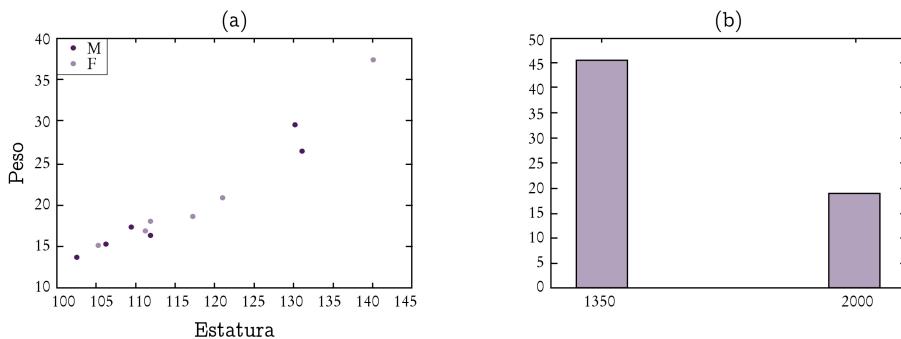
- (d) Hay que mirar la matriz fila por fila. Si, para cada alumno, miramos las diferencias entre la última y la primera prueba, vemos que los alumnos 6 y 7 tienen valores negativos.
- (e) Mejor crecimiento no significa mejor alumno, pues esto depende de qué nota partió. Vemos que el alumno 6, que tuvo un crecimiento negativo, tiene el mayor promedio. Ponga en un mismo gráfico las notas de los alumnos 6, 10 y 12. El mejor alumno podría ser el 6 o 12.
7. Los resultados son:
- La media en Fahrenheit es  $74^{\circ}$ , luego, la media en Celsius es:  $(74 - 32) / 1,8 = 23,33^{\circ}$ .
  - La temperatura promedio real en Fahrenheit es  $74^{\circ} + 5^{\circ} = 79^{\circ}$ . Luego, es  $(79 - 32) / 1,8 = 26,11^{\circ}$  Celcius.
8. (a) 100; (b) 10; (c) 8.
9. Los diagramas de caja se encuentran en la figura 9 (a) y (b), respectivamente.



10. Las respuestas son:
- Cada columna representa una variable con diferentes unidades de medición.
  - Medias: 7; 20,51; 116,54. Desviaciones estándares: 2,041; 6,75; 11,27. Los coeficientes de variación (desviación estándar/ media): 0,292; 0,329; 0,097. La estatura es la variable que varía menos.
  - El promedio y la desviación estándar serían divididos por 100. El coeficiente de variación no cambiaría.
  - Se observa, como era de esperar, que a mayor altura mayor peso (Figura (a)). Es conveniente separar las estadísticas por género. Calcule los promedios del peso y estatura por género y compare.

11. Las respuestas son:

- Construyendo el gráfico de barras de los promedios de recursos de agua en 1950 y 2000 se observa un claro descenso (Figura (b)).
- Ahora hay que trabajar por fila. Si denotamos  $X$  la disponibilidad de agua de una región, el porcentaje de descenso se calcula:  $100 \frac{X(2000) - X(1950)}{X(1950)}$ , obtenemos: 73,0%; 61,8%; 23,7%; 45,7%; 68,4%; 38,6%; 58,9%.
- El cambio climático, el crecimiento de las poblaciones, el aumento de consumo per capita, etc...



12. Los resultados son:

- La distribución de la edad de los hombres y mujeres en distintas fechas.
- Se ven dos fenómenos: una disminución de la natalidad y un aumento de la población mayor de edad.
- La esperanza de edad va a seguir creciendo. Si la natalidad no crece, los jóvenes no podrán seguir soportando a los mayores.

13. Los resultados son:

- Los tiempos de recorrido son:  $t_1 = 10/80 = 0,125$  (7'30'');  $t_2 = 50/100 = 0,5$  (30') y  $t_3 = 10/40 = 0,25$  (15'). Se demora en total 0,875 (52'30'').  $V = \frac{10+50+10}{0,875} = 80$  km/h.
- La media aritmética ponderada por la distancia :  $(10 \times 80 + 50 \times 100 + 10 \times 40)/70 = 88,6$  km/h.

14. Los resultados son:

- Si  $V$  es el volumen de la piscina, la potencia de la bomba a gas es:  $p_g = \frac{V}{4}$  y la de la bomba eléctrica es:  $p_e = \frac{V}{6}$ . Luego, el tiempo  $T$  para vaciar la piscina, usando las dos bombas al mismo tiempo, es tal que:  $Tp_g + Tp_e = V$ . Se deduce que

$$T = \frac{V}{p_g + p_e} = \frac{V}{V\left(\frac{1}{4} + \frac{1}{6}\right)} = \frac{1}{\frac{1}{4} + \frac{1}{6}} = \frac{4 \times 6}{4 + 6} = 2,4$$

Se tomará 2 horas y 24 minutos.

- (b) La media armónica de 4 y 6 es igual a  $\frac{2}{\frac{1}{4} + \frac{1}{6}} = 4,8$ , o sea, 4 horas y 48 minutos.

Observemos que es el doble del tiempo calculado anteriormente.

- (c) La media armónica de  $a$  y  $b$  es igual a  $H = \frac{2}{\frac{1}{a} + \frac{1}{b}} = \frac{2ab}{a+b}$

- (d) La media geométrica de  $a$  y  $b$  es:  $G = \sqrt{ab}$  y la media aritmética es;  $A = \frac{a+b}{2}$ .

Se deduce que la media armónica puede escribirse:  $H = \frac{G^2}{A}$  y, entonces  $G = \sqrt{AH}$ .

15. Los resultados son:

- (a) 0 2 2 4 4 5 5 6 6 7 8 8 8 9 9 9 10 y 0 2 2 4 4 5 5 6 6 7 8 8 8 9 9 9 9 10.

- (b) Serie 1: el promedio es 6 y la mediana es 6. Serie 2: el promedio es 6,17 y la mediana es 7.

- (c) Serie 1: ecuación 2.2: 7,88 y ecuación 2.3: 2,35 y Serie 2: ecuación 2.2: 7,92 y ecuación 2.3: 2,39.

- (d) Si tomamos el valor 6,5: Serie 1: ecuación 2.2: 8,13 y ecuación 2.3: 2,38 y Serie 2: ecuación 2.2: 8,03 y ecuación 2.3: 2,39. Observe que los valores respectivos son mayores que los obtenidos en el caso anterior. Repita con el valor 9. Compare.

- (e) Acabamos de ver que en el caso par no hay unicidad para la ecuación 2.3. Para el caso impar es única, lo es también para la ecuación 2.2.

16. Los resultados son:

- (a) Ver Figura (c).

- (b) 18 minutos.

- (c) Entre 15 y 22 minutos.

17. Los resultados son:

- (a) En Sector Norte el cuantil 55% es 182 y en el Sector Sur el tercer cuartil es 168.

- (b) Los coeficientes de variación son 0,137 y 0,158. Son bastante diferentes.

- (c) El nuevo promedio será 165,04 (148,41; 191,09; 123,19; 176,54; 152,29; 179,45; 184,30).

- (d) Figura (d). Los dos sectores tienen recorridos parecidos, pero el Norte muestra una fuerte asimetría con valores cargados a la derecha y un valor atípico bajo.

18. Los resultados son:

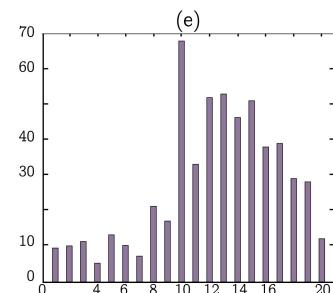
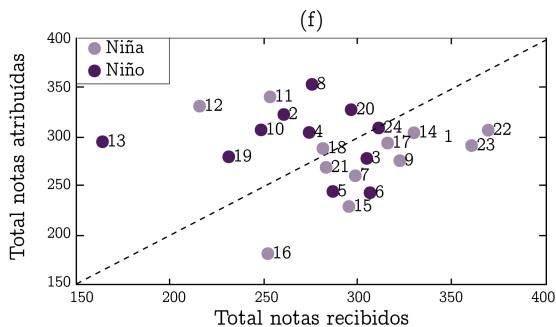
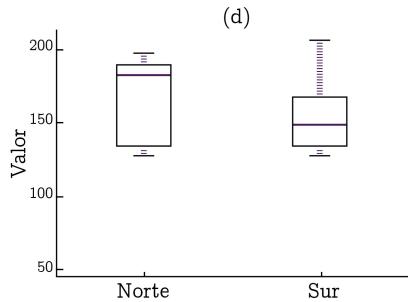
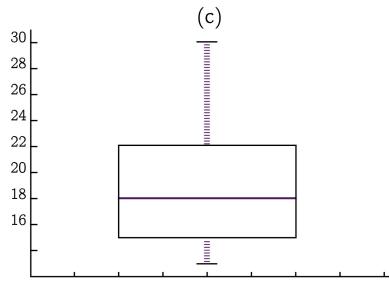
- (a) La nota atribuida más frecuente es 10, que es la nota promedio. Ver Figura (e).

- (b) Un alumno por debajo de la recta, tiene tendencia a recibir mejores notas que las que atribuya. Un alumno por arriba de la recta, tiene tendencia a atribuir mejores notas que las que recibe (Figura (f)).

- (c) Las niñas tienden a atribuir notas más altas que los niños.

- (d) Son niñas que atribuyen altas notas, pero no reciben las mejores notas.

- (e) El niño 13 es duro con sus compañeros, pero no recibe malas notas.  
 (f) Las niñas 12 y 16 atribuyen relativamente bajas notas. Sin embargo, la 16 atribuye las peores notas, pero la 12 recibe muy malas notas.



19. (a) Se construyó la tabla que muestra por región los porcentajes de alumnos que rindieron el SIMCE. Se obtiene dividiendo el número de alumnos de cada dependencia de una región por el total de alumnos de la región. El total da los porcentajes por dependencia sobre todos los alumnos. Se observa que, en la Región Metropolitana, la proporción de colegios municipales es menor en relación a las otras regiones. En la segunda región, la proporción de colegios subvencionadas es la menor.

Porcentajes de alumnos que rindieron el SIMCE

Región	Municipal	Part. pagado	Part. subvencionado	Total
1	42,6	2,1	55, 3	100
2	66,8	8,0	25,2	100
3	59,0	3,0	38,0	100
4	54,2	2,4	43,4	100
5	39,9	8,7	51,4	100
6	56,4	6,3	37,4	100
7	54,3	3,4	42,2	100
8	57,6	4,4	38,0	100
9	45,8	2,7	51,5	100
10	58,9	4,3	36,8	100
11	40,6	0	59,4	100
12	55,5	7,8	3,7	100
13	27,0	11,0	62,0	100
Total	42,8	7,2	50,0	100

- (b) En la segunda tabla se puede apreciar los puntajes promedio por dependencia (última fila), los puntajes por región (última columna) y por cruce de dependencia con región. Vemos promedios por región relativamente parecidos en comparación de las diferencias debidas a la dependencia. Esta diferencia en la dependencia es bastante estable de una región a otra.

### Capítulo 3

1. Lo anterior corresponde a (b) “un muestreo aleatorio simple”.
2. La solución es  $45 \times 1200/100 = 540$ .
3. La solución es  $5 * 3000/100 = 150$ .
4. Una muestra estará correctamente sacada cuando: (I) “ Sea aleatoria”, (III) “Sea a partir de toda la población” y (IV) “Posea un tamaño adecuado”. Luego, la respuesta es (d).
5. No estoy de acuerdo. Todas las secuencias de 6 números distintos del loto tienen la misma probabilidad de salir.
6. No se verifica que las mujeres a quienes se aplicó la encuesta tengan hijos, ni si asisten a las guarderías infantiles, lo que podría explicar el alto porcentaje de no respuesta. Además no se menciona el tipo de muestreo que se utilizó para seleccionar a las mujeres.
7.
  - (a) Si  $Z \sim \mathcal{N}(0, 1)$ ,  $\mathbb{P}(Z \leq -0,78) = 0,2177$  y  $\mathbb{P}(Z \geq -3,2) = 0,9993$ .
  - (b) Si  $X \sim \mathcal{N}(4, 2)$ ,  $\mathbb{P}(X \leq 3,1) = 0,3264$  y  $\mathbb{P}(X \geq 2,8) = 0,7257$ .
  - (c)  $u = 6,5631$ .

8. Definamos  $X = 1$  si la pareja tiene un hijo y  $X = 0$  si es hija.  $X$  sigue una distribución de Bernoulli  $\mathcal{B}(0,5)$  ( $\mathbb{P}(X = 1) = \mathbb{P}(X = 0) = 1/2$ ). Sea  $Y$  el número de hijos en 6 descendientes.  $Y \sim \text{Binomial}(6, 1/2)$ . Calculamos

$$\mathbb{P}(Y = 2) = \binom{6}{2} 0,5^2 0,5^4 = 15 \times 0,0156 = 0,2344$$

Se puede encontrar este valor en las tablas de distribución Binomial en Anexo, también se puede usar la aproximación normal.

9. El tamaño de la muestra debe ser 156. No requiere ser 300.

10. La proporción en la población que produce el mayor Error Cuadrático Medio es 0,50, que produce la mayor varianza de la distribución de Bernoulli:  $p \times (1 - p)$ .

11. (a) La distancia promedio: 12 km, con un error estándar de 0,6 km.

(b) Margen de error:  $2 \times 0,6 = 1,2$  (se approximó 1,96 a 2).

(c)  $[12 - 1,2; 12 + 1,2] = [10,8; 13,2]$ .

12. El error estándar máximo se obtiene con una proporción igual a 1/2:  $\frac{1}{2}\sqrt{\frac{1}{n}}$ . Para un nivel de confianza del 90 % se usa el coeficiente igual a 1,64 ( $\mathbb{P}(|z| \leq 1,64) = 0,90$ ). El producto de los dos números anteriores debe ser igual a 0,02, del cual se despeja la cota inferior para  $n = 1681$ .

13. El intervalo más ancho es el de nivel de confianza  $1 - \alpha = 99\%$ .

14. La solución es (a). En efecto, si la muestra es más grande, hay mayor posibilidad de encontrar un porcentaje de niñas cercano a 50 % en la población. Luego, es más probable de alejarse rápidamente del 50 % en una muestra pequeña que en una muestra grande. Pueden calcular  $\mathbb{P}(\hat{p} > 0,70)$  donde  $\hat{p}$  es la proporción muestral. Aproximando a la normal  $\hat{p} \sim \mathcal{N}(0,5; \frac{0,5}{\sqrt{n}})$  con  $n$  el tamaño de la muestra. Usando la  $\mathcal{N}(0,1)$ , esta probabilidad vale:  $\mathbb{P}(Z > \frac{0,70 - 0,50}{0,5/\sqrt{n}})$ , probabilidad que vale aproximadamente 0,10 para 10 niños y 0,000 para 100 niños.

15. La media  $\bar{x}$  del SIMCE tiene una distribución aproximada  $\bar{x} \sim \mathcal{N}(\mu, \sqrt{\frac{2400-300}{2399} \frac{\sigma}{\sqrt{300}}})$ ,

donde  $\sigma = 49$ . Sea  $s = \sqrt{\frac{2400-300}{2399} \frac{49}{\sqrt{300}}} = 2,647$ .

(a) Para 99 %, obtenemos  $[260 - 2,576 \times 2,647; 260 + 2,576 \times 2,647] = [253,2; 266,8]$ .

El valor real que es de 254,6 cae en el intervalo a 99 %.

(b) Para un nivel de confianza a 95 %, el intervalo es :  $[260 - 1,96 \times 2,647; 260 + 1,96 \times 2,647] = [254,8; 265,2]$ . El valor real que es de 254,6 no cae en el intervalo a 95 %

(c) En este caso se asume que la varianza de la media muestral es:  $\frac{49}{\sqrt{300}} = 2,829$ . Luego, obtenemos los intervalos  $[252,7; 267,3]$  y  $[254,5; 265,5]$  para 99 % y 95 %, respectivamente.

(d) Los intervalos de (c) son más grandes que los de (a) y (b). Conviene usar la corrección para población finita.

16. Los resultados son:

- (a) [ 40,65; 45,35].
- (b) n=138.

17. Los resultados son:

- (a) Bajo  $H_o$ , la media  $\bar{x} \sim \mathcal{N}(80, 7/\sqrt{200})$ . Como la hipótesis alternativa es menor que la nula, la región crítica tiene la forma  $\mathcal{R} = \{\bar{x} \leq a\}$  y  $\mathbb{P}(\bar{x} \leq 79, 186) = 5\%$ . Se decide que el aeropuerto tiene la razón. Tomando  $\bar{x} \sim \mathcal{N}(78, 7/\sqrt{200})$  el otro error es  $\beta = \mathbb{P}(\bar{x} > 79, 186) = 1\%$ .
- (b) Bajo  $H_o$ , la media  $\bar{x} \sim \mathcal{N}(78, 7/\sqrt{200})$ . Como la hipótesis alternativa es mayor que la nula, la región crítica tiene la forma  $\mathcal{R} = \{\bar{x} \geq a\}$  y  $\mathbb{P}(\bar{x} \geq 78, 814) = 5\%$ . Se decide que los vecinos tienen la razón. Tomando  $\bar{x} \sim \mathcal{N}(80, 7/\sqrt{200})$  el otro error es  $\beta = \mathbb{P}(\bar{x} < 78, 814) = 1\%$ .
- (c) Tuvimos conclusiones diferentes en los casos (a) y (b). El aeropuerto quiere que se concluya que  $\mu = 78$  y que se haga con el mínimo error. Le conviene el caso (a).

18. Los resultados son:

- (a) El fabricante asegura que, con 4 litros de bencina, la máquina puede funcionar al menos 300 horas (5 horas) y se pone en duda esta eficiencia cuando se tenga menos de 300 minutos. Si llamamos  $\mu$  la verdadera media, las hipótesis nula y alternativa son:  $H_o : \mu = 300$ , contra  $H_1 : \mu < 300$ .
- (b) La media encontrada en la muestra es 294 minutos y la desviación estándar es 24 minutos en la población. Bajo  $H_o$  la media muestral  $\bar{x} \sim \mathcal{N}(300, 24/\sqrt{85})$ , o sea  $\bar{x} \sim \mathcal{N}(300; 2, 603)$ . Si  $Z = \frac{\bar{x}-300}{2,603}$ ,  $Z \sim \mathcal{N}(0, 1)$  y  $\mathbb{P}(Z \leq -1, 645) = 0, 05$ . Luego,  $\mathbb{P}(\bar{x} \leq 295, 72) = 0, 05$ . La media de la muestra, que es 294 minutos, se encuentra en la región crítica que es  $\mathcal{R} = \{\bar{x} \leq 295, 72\}$ . Se puede rechazar  $H_o$  para un nivel de significación de 5%.
- (c) El p-valor se obtiene de  $\mathbb{P}(\bar{x} \leq 294) = \mathbb{P}(Z \leq \frac{294-300}{2,603}) = \mathbb{P}(Z \leq -2, 305) = 0, 0106$ . Se concluye que se puede rechazar  $H_o$ , con un error de tipo I al menos igual a 1,06 %.

19. Denotamos  $\bar{x}$  la media muestral y  $\mu$  la media de la población y  $Z \sim \mathcal{N}(0, 1)$ .

- (a)  $H_o : \mu = 22$  y  $H_1 : \mu < 22$ .
- (b)  $\bar{x} \sim \mathcal{N}(\mu, 0, 3/\sqrt{15})$ . Calculamos el p-valor:

$$\mathbb{P}(\bar{x} \leq 21, 4) = \mathbb{P}(Z \leq (21, 4 - 22)/(0, 3/\sqrt{15})) = \mathbb{P}(Z \leq -7, 746) = 0, 000.$$

La queja del consumidor se justifica: se rechaza  $H_o$  para un nivel de significación casi nulo.

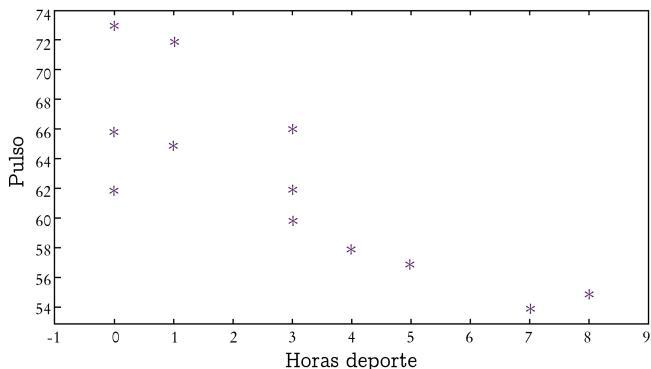
20. Se esperaría que los peces pequeños sean más afectados.

- (a) Sean  $\mu_1$ ,  $\sigma_1$  y  $n_1$  la media, desviación estándar y n° de peces, para la radioactividad de las escamas o de los ojos, en los peces grandes y  $\mu_2$ ,  $\sigma_2$  y  $n_2$  para los peces pequeños. Tenemos  $H_0 : \mu_2 - \mu_1 = 0$ , contra  $H_1 : \mu_2 - \mu_1 > 0$ . Bajo  $H_0$  la media muestral tiene una distribución aproximada:  $\bar{x}_1 - \bar{x}_2 \sim \mathcal{N}(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$ .
- (b) Para la radioactividad de las escamas:  $s = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 17,97$ . Encontramos una región crítica para 5%:  $\mathcal{R} = \{\bar{x}_1 - \bar{x}_2 > 30,56\}$ . La diferencia  $\bar{x}_1 - \bar{x}_2 = 55,9$  que cae en la región crítica. Se concluye que la radioactividad de las escamas es mayor en los peces pequeños. El p-valor es igual a: 0,0009. Este valor es muy pequeño, por lo que podemos rechazar sin mucho riesgo.
- (c)  $s = 2,64$ . La diferencia  $\bar{x}_1 - \bar{x}_2 = 2,3$ . Encontramos una región crítica para 5%:  $\mathcal{R} = \{\bar{x}_1 - \bar{x}_2 > 5,34\}$ . El p-valor es igual a 0,19. El p-valor es alto y la diferencia no cae en la región crítica. No se rechaza  $H_0$  para la radioactividad de los ojos.

#### Capítulo 4

1.

- (a) Ver Figura adjunta
- (b) Se observa una tendencia de decrecimiento del pulso cuando aumenta el tiempo de deporte.
- (c)  $-1,85 \times 2 + 68,1 = 64,4$ .
- (d)  $-1,85 \times 4 + 68,1 = 60,7$ . Se planteo un modelo lineal. Es razonable en un intervalo de tiempo, pero el pulso no puede seguir disminuyendo linealmente cuando aumenta el tiempo de dedicado al deporte. Se espera una función asintótica.



2.

- (a) Se desprende de la sección 4.6.
- (b)  $\hat{a}$  es proporcional a  $r_{X,Y}$  y el coeficiente de proporcionalidad depende de las varianzas de las dos variables. Si las dos varianzas son iguales a 1, entonces  $\hat{a} = r_{X,Y}$ .

3.

- (a) El modelo se escribe:

$$\text{Venta} = a \times \text{gasto} + b + e = -61,491 + 10,949 \times \text{gasto} + e$$

La pendiente de la recta de regresión es positiva. Cuando se aumenta el gasto en publicidad, aumenta la venta.

- (b) Los residuos son: 2.00364 -2.74016 14.41356 42.51604 -12.22777 -21.71537 -85.94678 -40.69058 54.56561 49.82181. La suma es nula.
- (c) La predicción es  $-61,491 + 10,949 \times 550 = 5960,5$ .
- (d) La recta de regresión que pasa por el origen es:  $\text{Venta} = c \times \text{gasto}$ . La solución de los mínimos cuadrados es:  $\hat{c} = \frac{\sum x_i y_i}{\sum x_i^2} = 9,613$ . La media de los residuos es igual a  $-143,35344$ .

4.

- (a) Porque el criterio de los mínimos cuadrados es el más utilizado. Sin embargo, se podría minimizar, por ejemplo, la suma de los valores absolutos de los errores. No es correcto.
- (b) Se calcula el coeficiente de correlación lineal y la varianza de  $X$ . Deducimos  $\hat{a}$  y  $\hat{b}$ . Se hacen los gráficos de dispersión para validar el modelo.
- (c) Hay un punto atípico que desvía la recta de regresión. El valor  $X = 50$  está fuera de la masa de los puntos no atípicos, por lo cual, tampoco la predicción sería buena.

5.

- (a)  $Y = -0,467X + 12,267$ . La pendiente es negativa: a mayor fuerza laboral agrícola, menor ingreso per capita.
- (b) La suma de los residuos al cuadrado es: 26,93, luego,  $\text{Var}(\hat{e}) = \frac{1}{15} \sum_{i=1}^{15} \hat{e}_i^2 = 1,80$ .
- (c) Los gráficos de validación y el coeficiente de correlación igual a  $-0,572$ , muestran un modelo muy deficiente. No se puede usar.
- (d) Por lo que vimos el nuevo modelo se escribe:  $Y = -46,7X + 1226,7$ .

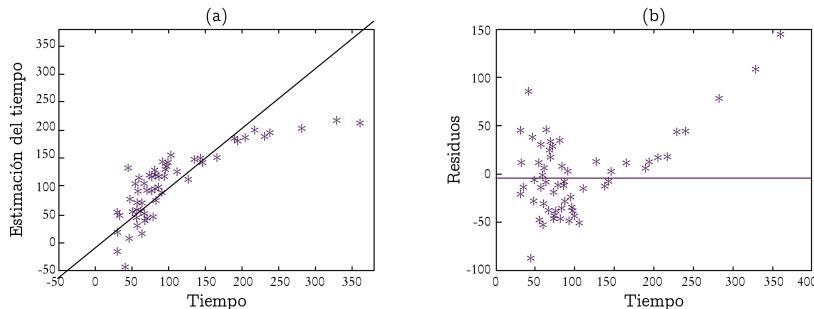
6.

- (a) Si bien el coeficiente de correlación lineal de 0,82 es bastante aceptable, el diagrama de dispersión del examen del hígado, con el tiempo de sobrevivencia del grupo de los 60 pacientes del médico, no muestra una relación muy lineal.

No se recomendaría al médico calcular sus predicciones a partir de la regresión lineal del tiempo sobre el examen del hígado.

- (b) La recta de regresión es:  $Y = 52,852X - 40,347$ . Los gráficos de la Figura 4.9 muestran que el modelo no es adecuado. En el gráfico (a) los puntos no son cercanos a la primera bisectriz. En el gráfico (b) no deberían mostrar tendencias. ¡Incluso se obtiene predicciones del tiempo negativas!
- (c) El gráfico de dispersión sugiere esta relación.
- (d) La nueva recta de regresión es:  $Y = 0,448X + 3,239$ . El coeficiente de correlación entre el logaritmo del tiempo y el examen de hígado es mayor que el anterior: 0,829.
- (e) Los gráficos de la Figura 4.10 son bastante aceptables para decir que tenemos un buen modelo, además de un coeficiente de correlación muy razonable.
- (f) Si  $X = 3,5$ , con el segundo modelo obtenemos el logaritmo del tiempo  $Y = 0,448 \times 3,5 + 3,239 = 4,807$ . Luego, el tiempo estimado es  $e^{4,807} = 122,36$ , o sea, un pronóstico cercano a los 122 meses.

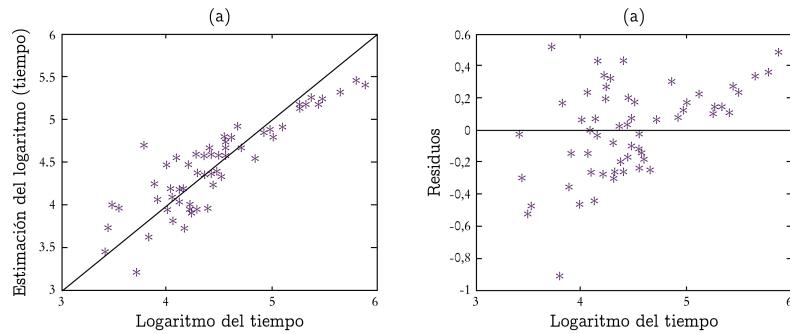
FIGURA 4.9. Recta de regresión y residuos



7.

- (a)  $\hat{b}$  existe y es único si  $\hat{a}$  existe.  $\hat{a}$  existe si  $\text{Var}(X) \neq 0$ , caso sin interés. Si  $\hat{a}$  existe, es único.
- (b) Permite ver si existe una tendencia lineal entre las dos variables y si algunos puntos son atípicos, en el sentido que se alejan del resto de los datos, y por lo tanto, no tienen el mismo comportamiento. En este caso hay que aislar los puntos atípicos.

FIGURA 4.10. Recta de regresión y residuos después de transformación logarítmica





## Anexo 2: Tablas Estadísticas

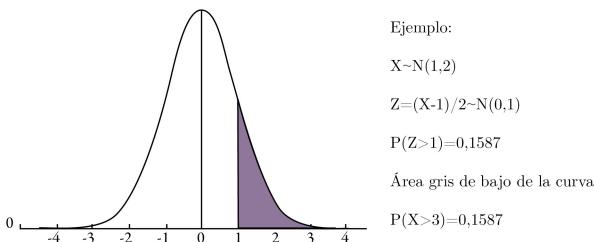


La tabla de distribución Normal  $\mathcal{N}(0, 1)$  no es fácil de usar. Tomando en cuenta que la distribución  $\mathcal{N}(0, 1)$  es simétrica con respecto de 0, la tabla presenta solamente las probabilidades de los valores positivos. Por ejemplo, para  $\mathbb{P}(Z \geq 1,03)$ , se busca la parte entera y la decimal de 1,03 que son, respectivamente, 1,0 en la primera columna y se busca la centesimal que es 0,03 en la primera fila. El valor de la probabilidad se encuentra, entonces, al cruce de la fila en el valor 1,0 y de la columna en el valor 0,03. La probabilidad es 0,1515.

Si se quiere  $\mathbb{P}(Z \leq -1)$ , basta tomar  $1 - \mathbb{P}(Z \geq 1) = 0,159$ .

Si  $X \sim \mathcal{N}(2, 3)$  y se quiere  $\mathbb{P}(X \leq 2,5)$  se normaliza  $X$ :  $Z = \frac{X-2}{\sqrt{3}}$  y se calcula  $\mathbb{P}(Z \leq \frac{2,5-2}{\sqrt{3}}) = \mathbb{P}(Z \leq 0,17) = 1 - \mathbb{P}(Z \geq 0,167)$ . En la tabla  $\mathbb{P}(Z \geq 0,17) = 0,4325$ . Se deduce  $\mathbb{P}(X \leq 2,5) = 0,5675$ .

## DISTRIBUCIÓN NORMAL



Decimal	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010

## DISTRIBUCIÓN BINOMIAL

<i>n</i>	<i>k</i>	0.05	0.10	0.15	0.20	<i>p</i>	0.25	0.30	0.35	0.40	0.45	0.50
1	0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000	0.5000
	1	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000	0.5000
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500	0.2500
	1	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000	0.5000
2	2	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500	0.2500
	3	0.001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250	0.1250
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250	0.1250
	1	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750	0.3750
3	2	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750	0.3750
	3	0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250	0.1250
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625	0.0625
	1	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500	0.2500
4	2	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750	0.3750
	3	0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500	0.2500
4	4	0.0000	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625	0.0625
	5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
5	1	0.2036	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563	0.1563
	2	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125	0.3125
5	3	0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125	0.3125
	4	0.0000	0.0005	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563	0.1563
5	5	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313	0.0313
	6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
6	1	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938	0.0938
	2	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344	0.2344
6	3	0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125	0.3125
	4	0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344	0.2344
6	5	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938	0.0938
	6	6	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078	0.0078
	1	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547	0.0547
7	2	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1641
	3	0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734	0.2734
7	4	0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734	0.2734
	5	0.0000	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641	0.1641
7	6	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547	0.0547
	7	7	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039	0.0039
	1	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313	0.0313
8	2	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094	0.1094
	3	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188	0.2188
8	4	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734	0.2734
	5	0.0000	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188	0.2188
8	6	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094	0.1094
	7	0.0000	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313	0.0313
8	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039	0.0039



## Bibliografía



- [1] Aliaga M., Gunderson B., *Interactive Statistics*, Prentice Hall, 2002.
- [2] Batanero C., Godino J., *Ánalisis de datos y su didáctica*, Universidad de Granada, 2001
- [3] Batanero C., *Didáctica de la Estadística*, Universidad de Granada, 2001.
- [4] Batanero C., *Recursos para la educación estadística en Internet*, UNO, 1998, 15,13-25.
- [5] Brook R. et al., *The Fascination of Statistics*, Marcel Dekker, 1986.
- [6] Cerda L., Lacourly N., *Apunte de Estadística, Estadía de Especialización en Matemática del MINEDUC*, 2004.
- [7] Goon A., Gupta M. & B. Dasgupta, *Fundamentals of Statistics*, The World Press Private, Calcutta, 1987.
- [8] Lacourly N., Soto J., *¿Qué dicen los datos? Introducción a la Estadística*, Notas para un curso PPF Del MINEDUC, 2002.
- [9] Moore D., McCabe G., *Introduction to the Practice of Statistics*, 3ra Edición, W H Freeman & Co, 1998.
- [10] Naiman A., Rosenberg R. & Zirkel G., *Understanding Statistics*, Mc Graw-Hill, 1996.
- [11] Newman J., *The World of Mathematics*, Simon & Schuster, New York, 1956.
- [12] Romagnoli P., *Probabilidades Doctas con discos, bolitas y urnas*, J.C Sáez Editores, Santiago, 2011
- [13] Ycart B., Curso por Internet, <http://1jk.imag.fr/membres/Bernard.Ycart/emel/index.html>
- [14] Yule G. U. *An Introduction to the theory of statistics*, C. Griffin, Londres, 1922.



## Índice de figuras



2.1. Diagrama de barras N° hermanos	34
2.2. Diagramas de barras	35
2.3. Diagramas de barras frecuencias acumuladas	36
2.4. Histogramas de las notas de la primera prueba de matemáticas	37
2.5. Histogramas del promedio SIMCE 2do medio 2006 por colegio	39
2.6. Histogramas del SIMCE con 8 y 23 intervalos	39
2.7. Histogramas SIMCE horizontal y vertical	40
2.8. Histogramas de la PSU matemática 2005	41
2.9. Histograma de una distribución simétrica	41
2.10. Diagramas de barras de la conducta y del género	42
2.11. Diagramas circulares de la conducta y del género	42
2.12. Un gráfico de tallos y hojas	43
2.13. Notas dos primeras pruebas	47
2.14. Notas de las dos primeras pruebas	48
2.15. Notas de la primera prueba	50
2.16. Notas de otra prueba	50
2.17. Distribución bimodal	51
2.18. Distribuciones SIMCE y PSU	52
2.19. Porcentajes acumulados y cuantiles	53
2.20. Cuartiles notas	54
2.21. Quintiles SIMCE 2006	55
2.22. Un gráfico de caja	55
2.23. Diagrama de cajas	56
2.24. Gráficos de cajas SIMCE	57
2.25. Histogramas SIMCE	57
2.26. Frecuencias alumnos SIMCE según NSE y dependencia	59
2.27. Promedio SIMCE según NSE y dependencia	60
2.28. Un termómetro	63
2.29. Pirámides de edades	66
2.30. Socio-matriz	68
3.1. Tamaño de la muestra y errores muestrales y no muestrales	73
3.2. Experimento con un jarro	76

3.3. Esquema de razonamiento en Estadística	79
3.4. Distribución de frecuencias de la media muestral	80
3.5. Error estándar de la media y tamaño de la muestra	84
3.6. La campana de Gauss	86
3.7. Distribución $\mathcal{N}(\mu, \sigma)$	87
3.8. Intervalos de confianza de 50 muestras	92
3.9. Niveles de confianza	92
3.10. Error de Tipo II	99
3.11. Errores de Tipo I y II	100
3.12. Región crítica del test	103
3.13. Diagrama de cajas	105
4.1. Ejemplos de los países	115
4.2. Signos de $(X - M_X)(Y - M_Y)$	116
4.3. Gráficos de dispersión y coeficientes de correlación	118
4.4. Errores del modelo	122
4.5. Gráficos para validez del modelo	125
4.6. Las dos rectas de regresión	127
4.7. Recta de regresión y residuos	131
4.8. Curva correcta	134
4.9. Recta de regresión y residuos	144
4.10. Recta de regresión y residuos después de transformación logarítmica	145

## Índice de nombres propios



- Bernoulli, 22  
de Moivre, 84  
Fisher, Sir Ronald, 95  
Galton, 85, 120,  
máquina de, 89  
Gauss, 84, 121  
Graunt, 21, 29  
Halley, 21  
Huygens, 22, 29  
Jouffret, 85  
Laplace, 22, 85  
Legendre, 84, 121  
Lippman, 85  
Pearson, 17, 115  
Quetelet, 44  
Simpson, 85  
Yule, 46



## Índice de Términos



- Azar  
juegos de, 21
- Coeficiente  
de correlación lineal, 115, 117, 124  
de correlación múltiple, 124  
de variación, 51
- Covarianza, 116
- Criterio de los mínimos cuadrados, 120
- Cuantil, 53
- Cuartil, 54
- Diagrama  
circular, 39  
de barra, 37, 40  
de caja, 52
- Distribución  
de frecuencias, 33, 37
- Distribución Normal, 84
- Error  
de Tipo I, 96, 98  
de Tipo II, 96, 98  
muestral, 72  
no muestral, 72
- Estadígrafos, 51
- Estadística, 22  
Descriptiva, 23, 25, 29  
Inferencial 24, 25, 72
- Estadísticas, 23
- Estimación  
por intervalo, 91
- Estimador, 79
- Función de densidad, 85
- Gráfico de tallo y hojas, 41
- Hipótesis  
bilateral, 102  
paramétrica, 97  
unilateral, 102
- Histograma, 37
- Intervalo intercuartílico, 54
- Método de los Mínimos Cuadrados, 85
- Media  
aritmética, 44, 50  
armónica, 45  
cuadrática, 45  
geométrica 45
- Mediana, 49
- Medida  
de dispersión, 47  
de posición, 43, 45
- Muestra, 71
- Muestreo  
aleatorio simple, 74  
Diseño, 73  
estratificado, 76  
por conglomerados, 77  
sistématico, 77  
Teoría, 73
- p-valor, 98
- Población, 31
- Quintil, 54
- Recorrido, 48
- Región crítica, 98
- Regresión  
Errores del modelo, 121  
lineal simple, 119  
predicciones, 124  
residuos, 123  
validación, 124
- Tablas de mortalidad, 21
- Test  
para dos medias, 105  
para una media, 100

para una proporción, 104  
Variable  
  continua, 33, 36, 37  
  discreta, 33, 37  
nominal, 33  
ordinal, 33  
Varianza, 48