ESTADÍSTICA PARA TODOS

Estrategias de pensamiento y herramientas para la solución de problemas

Dra. Diana M. Kelmansky

Colección: LAS CIENCIAS NATURALES Y LA MATEMÁTICA

Colección: LAS CIENCIAS NATURALES Y LA MATEMÁTICA

ESTADÍSTICA PARA TODOS

Estrategias de pensamiento y herramientas para la solución de problemas

Dra. Diana M. Kelmansky

autoridades

Presidente de la Nación
Dra. Cristina Fernández de Kirchner

Ministro de Educación
Dr. Alberto E. Sileoni

Secretaria de Educación
Prof. María Inés Abrile de Vollmer

Directora Ejecutiva del Instituto Nacional de Educación Tecnológica
Lic. María Rosa Almandoz

Director Nacional del Centro Nacional de Educación Tecnológica
Lic. Juan Manuel Kirschenbaum

Director Nacional de Educación Técnico Profesional y Ocupacional
Ing. Roberto Díaz

Ministerio de Educación.
Instituto Nacional de Educación Tecnológica.
Saavedra 789. C1229ACE.
Ciudad Autónoma de Buenos Aires.
República Argentina.
2009

ESTADÍSTICA PARA TODOS

Estrategias de pensamiento y herramientas para la solución de problemas

Dra. Diana M. Kelmansky

Colección: LAS CIENCIAS NATURALES Y LA MATEMÁTICA

Colección “Las Ciencias Naturales y la Matemática”.
Director de la Colección: Juan Manuel Kirschenbaum
Coordinadora general de la Colección: Haydeé Noceti.

Industria Argentina

ISBN 978-950-00-0713-9

Director de la Colección:
Lic. Juan Manuel Kirschenbaum

Coordinadora general y académica de la Colección:
Prof. Ing. Haydeé Noceti

Diseño didáctico y corrección de estilo:
Lic. María Inés Narvaja
Ing. Alejandra Santos

Coordinación y producción gráfica:
Tomás Ahumada

Diseño gráfico:
Martin Alejandro Gonzalez

Ilustraciones:
Diego Gonzalo Ferreyro

Retoques fotográficos:
Roberto Sobrado

Diseño de tapa:
Tomás Ahumada

Administración:
Cristina Caratozzolo
Néstor Hergenrether

Nuestro agradecimiento al personal del Centro Nacional de Educación Tecnológica por su colaboración.

Kelmansky, Diana
Estadística para todos / Diana Kelmansky; dirigido por Juan Manuel Kirschenbaum.
- 1a ed. - Buenos Aires: Ministerio de Educación de la Nación. Instituto Nacional de Educación Tecnológica, 2009.
272 p. ; 24x19 cm. (Las ciencias naturales y la matemática / Juan Manuel Kirschenbaum.)

ISBN 978-950-00-0713-9

1. Estadística.
I. Kirschenbaum, Juan Manuel, dir.
II. Título

CDD 310.4

Fecha de catalogación: 21/08/2009

Impreso en Artes Gráficas Rioplatense S. A., Corrales 1393 (C1437GLE), Buenos Aires, Argentina.

Tirada de esta edición: 100.000 ejemplares

Dra. Diana M. Kelmansky

La Autora

Diana M. Kelmansky es Doctora en Matemática de la Universidad de Buenos Aires (UBA-1991).
Actualmente se desempeña como Profesora Adjunta en el Instituto de Cálculo de la Facultad de Ciencias Exactas y Naturales (UBA) y como Vicedirectora de la Carrera de Especialización de Estadística para Ciencias de la Salud.
Se desempeñó desde 1992 hasta 1994 como consultora del Instituto Nacional de Estadísticas y Censos (INDEC), en el marco del Programa de las Naciones Unidas para el Desarrollo (PNUD).
Desde el año 2002 hasta el 2004 fue consultora invitada por la Organización Mundial de la Salud (OMS) en el Plan de Análisis Estadístico sobre Crecimiento y Desarrollo.
Es Embajadora Educativa de la Sociedad Americana de Estadística (ASA) desde el 2005.
Dictó cursos y conferencias sobre microarrays en Argentina, México, Chile y España.
Ha publicado trabajos en revistas especializadas, nacionales e internacionales, y ha dictado numerosos cursos y conferencias en congresos abordando temáticas referidas, tanto a la estadística teórica como a sus aplicaciones en biología, economía y medicina.

ÍNDICE

1. Introducción

La **estadística** puede ser **divertida**, fácil y también **útil**.

Se la utiliza todos los días:

- Para justificar apuestas sobre el resultado de un partido de fútbol los simpatizantes comparan los rendimientos de los equipos utilizando, por ejemplo, los porcentajes de partidos ganados como local y como visitante.
- Durante la transmisión de un partido de tenis por televisión, los relatores cuentan la cantidad de tiros ganadores, puntos de quiebre aprovechados, errores no forzados, saques ganadores.
- Para diseñar pautas publicitarias, los publicistas consultan la planilla diaria de ratings (radio o televisión).
- En un mercado los consumidores observan cómo se distribuyen los precios entre los distintos puestos para realizar la mejor compra que combine calidad y precio.
- Para decidir qué alumna/o será abanderada/o de la escuela, el/la directora/a compara las notas de todos los alumnos del último año y elige el mejor promedio.

La necesitan:

- Los profesionales de la salud, para entender los resultados de las investigaciones médicas.
- Los economistas, porque cálculos eficientes les permitirán llegar al fondo de la cuestión que analizan.
- Los docentes cuando se enfrentan al problema de evaluar el rendimiento de los alumnos.
- Los sociólogos para diseñar y procesar sus encuestas.
- Los responsables de la calidad en un proceso productivo, al detectar las piezas defectuosas y controlar los factores que influyen en la producción de las mismas.
- La industria farmacéutica para desarrollar nuevos medicamentos y establecer las dosis terapéuticas.
- Los ciudadanos, para sacar sus propias conclusiones sobre los resultados de las encuestas políticas, los índices de precios y desocupación, y los resultados estadísticos que habitualmente se presentan en los medios masivos de comunicación (diarios, revistas, radio, televisión).

Muchas veces, las noticias surgen luego de varias etapas de elaboración. Sus primeros protagonistas son encuestadores, investigadores de mercado, médicos, técnicos gubernamentales y científicos de universidades o institutos. Ellos son la fuente original de la información estadística; publican sus resultados en revistas especializadas o en comunicados de prensa.

A partir de allí, entra en juego el segundo grupo: los periodistas, que pueden estar más apurados, a la caza de resultados que les permitan obtener un titular.

Finalmente, hay un tercer grupo: el de los consumidores de la información, o sea todos nosotros. Estamos frente al desafío de escuchar, leer, ver y decidir respecto a ella.

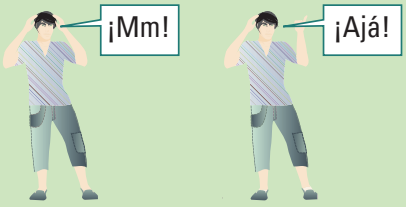
Los métodos estadísticos forman parte de cada paso de una buena investigación, desde el **diseño** del estudio, la **recolección** de los datos, la **organización** y el **resumen** de la información, el **análisis**, la elaboración de las **conclusiones**, la discusión de las limitaciones y, por último, el diseño de un **próximo estudio** a fin de dar respuesta a las nuevas preguntas que pudieran surgir.

En cualquiera de las etapas de este proceso puede haber errores. Pueden, o no, ser intencionales. **Es posible mentir con estadísticas, pero es mucho más fácil mentir sin estadísticas.**

Proponemos construir herramientas que permitan:

- Descubrir resultados engañosos.
- Obtener buenos datos.
- Distinguir entre lo que se puede y no se puede concluir a partir de una muestra.
- Entender tablas y gráficos.
- Comprender el significado de margen de error.
- Construir e interpretar intervalos de confianza.
- Tomar decisiones en base a los datos.
- Llevar a cabo estudios estadísticos sencillos

Este libro se estructura en base a ejemplos. Algunos de ellos reaparecen en capítulos sucesivos con una profundidad creciente poniendo el énfasis en el desarrollo de los conceptos. Para entenderlos y aprehenderlos hace falta **pensar**. Habrá párrafos que requerirán de varias lecturas, hasta que... "¡Eureka!", se comprende su significado.



Todos los cálculos presentados, tanto en el texto como en los ejercicios utilizan operaciones aritméticas simples, realizables con una calculadora. Debe tomarse **tiempo para pensar** las respuestas a los ejercicios sin mirar las soluciones. Estas son únicamente una guía, y para su verificación. Aunque algunas explicaciones y detalles no se dan en las soluciones estas deben formar parte de las respuestas completas a los mismos.

Utilizaremos la palabra **estadístico/a** con **cuatro significados diferentes** que, según el contexto, será fácil distinguir:

1. La **estadística como disciplina de estudio**. Siempre estará en **singular**.

2. La **estadística** o las **estadísticas** como resultados que presentan organismos de estadística oficiales como, por ejemplo, la Dirección de Estadísticas e Información de Salud -DEIS- del Ministerio de Salud y Ambiente de la Nación (http://www.deis.gov.ar/CapacitacionFetal/sistema.htm).

3. Un **estadístico** como un **procedimiento** para **obtener un número** a partir de valores de una encuesta.

4. Un **estadístico** o una **estadística** como una **persona** que tiene a la estadística como **profesión**.

De aquí que, cuando hablemos de los estadísticos o las estadísticas tendremos que ver si se trata del plural de 2, 3 ó 4.

2. Un poco de historia

Desde el antiguo Egipto hasta las actuales computadoras, pasando por los adelantados de la astronomía y del estudio de la herencia.

La **historia antigua** de la **estadística** se remonta al registro de la población que hicieron los egipcios, hebreos, chinos, griegos y romanos, desde hace unos 20 a 50 siglos. Se trata de mediciones que ya realizaba el **estado** con fines tributarios y de enrolamiento militar.

Sin embargo, las **ideas** y **herramientas estadísticas** son más recientes, surgieron lentamente de las dificultades que se plantean al trabajar con datos.

Hace dos siglos, los investigadores ya enfrentaban el problema de obtener diferentes valores **para un mismo concepto**. Debían **combinar los resultados de muchas observaciones** que, por más que las realizaran con extremo cuidado, no coincidían. Tal como sigue ocurriendo actualmente cuando, por ejemplo, se mide la altura de un niño varias veces. Se trata de la **variabilidad** debida a los **errores de medición**, cuando se obtienen resultados diferentes al medir lo mismo más de una vez.

Otro tipo de variabilidad surge, por ejemplo, en el caso de individuos de una misma población que, respecto a una misma característica, son diferentes entre sí. Por ejemplo, diferentes niños de la misma edad y género tienen distintas estaturas.

En la Argentina: El primer censo nacional se realizó en 1869. Diversos organismos tuvieron a su cargo la producción de estadísticas oficiales hasta la creación del Instituto Nacional de Estadísticas y Censos (INDEC) en 1968 (http://www.indec.gov.ar/indec/indec_historia.asp).

La Sociedad Argentina de Estadística (SAE) fue creada en 1952; es una organización técnico científica, sin fines de lucro, dedicada a promover el desarrollo de la Estadística en nuestro país.

Hacia comienzos del siglo XIX, los **astrónomos** utilizaban en forma generalizada **métodos estadísticos** y escribían **textos** razonablemente **sencillos** para explicarlos. Para describir la variabilidad de sus observaciones, resultantes de los inevitables errores de medición, utilizaban como modelo matemático la distribución Normal o Gaussiana, porque les permitía explicarla con solo dos valores: la media y el desvío (Se verá con más detalle en los Capítulos 17, 18 y 20).

La distribución Normal también se utilizó para caracterizar la variabilidad entre individuos de una población, respecto de alguna característica, como por ejemplo, el perímetro cefálico. Ya no se trata de una **variabilidad** debida a los errores de medición, sino a las **diferencias entre un individuo y otro.**

Originalmente, la **estadística** estuvo limitada al cálculo de **medidas resumen.** Por esa razón, existe una directa asociación entre "hacer una estadística" y "calcular un promedio o un porcentaje". Esta última, es la estadística que encontramos habitualmente en los medios de comunicación: **promedios, porcentajes, gráficos de barras.**

Pero la estadística es más que el cálculo de promedios y porcentajes. Específicamente, se trata de hallar el rango de valores dentro del cual pueden encontrarse los datos o la mayoría de ellos, es decir, caracterizar su **variabilidad** y, más generalmente, su **distribución** completa. Conocer la distribución de los datos es importante. Un **promedio** puede tener **significados muy diferentes** según sea la forma en que se distribuyen los valores.

Algunos realizaron estimaciones con singular éxito: entre julio y septiembre de 1882, sí, ¡1882!, el astrónomo y matemático, canadiense/norteamericano, Simon Newcomb logró una estimación bastante precisa (299.810 km/s) de la velocidad de la luz realizando una combinación ponderada de una serie de observaciones. Actualmente, se considera que la velocidad de la luz en el vacío es 299.792,458 km/s, o sea aproximadamente 300 mil km/s.

Otros, en cambio, sin éxito: en 1869, el economista inglés, William Jevons, fue acusado de combinar mal los precios de diferentes productos en un índice para estudiar las variaciones del precio del oro. Los índices de precios fueron y seguirán siendo siempre un gran dolor de cabeza.

Cuando se quiere obtener conclusiones respecto de **toda la población** pero no es posible, o no es deseable, registrar datos de esa población completa, se los obtiene de algún subgrupo o **muestra de la población.** Este proceso se denomina **inferencia estadística.**

La **inferencia estadística** como disciplina nació en la primera mitad del Siglo XX con el surgimiento de los **diseños estadísticos** para obtener datos y el desarrollo de métodos para analizarlos. Sin embargo, fueron los últimos 30 años, en especial con el advenimiento de las computadoras, los que vieron la explosión de su desarrollo y aplicación.

La invasión del **mundo digital** a nuestras vidas (computadoras, acceso a Internet, teléfonos celulares, cámaras digitales) acelera los procesos de obtención y difusión de la información. Todos los campos de estudio ponen **mayor énfasis en los datos.**

La estadística se ha transformado en un método central del conocimiento. Toda **persona educada** debería estar familiarizada con los **conceptos estadísticos.**

3. Los datos son noticia

Cada mañana nos enfrentamos con una gran cantidad de información estadística que abarca prácticamente todo: desde deportes, política y economía, hasta los avisos publicitarios.

Presentaremos varios ejemplos reales que ilustran esta situación, e iniciaremos una línea de análisis, que ampliaremos en los capítulos siguientes, para determinar si las conclusiones que presentan son las adecuadas.

□ 3.1 Encuestas de opinión

Una encuesta de opinión es un mecanismo para acercarse a la visión que tiene el conjunto de la sociedad, o algún subgrupo, sobre un determinado tema. Se utilizan diferentes métodos: preguntas en la calle, por teléfono, mediante citas previas, etc. Generalmente, requieren respuesta voluntaria.

Ejemplo: Un diario presenta los resultados de una encuesta de opinión.

Se había consultado a la gente si los **mensajes del correo electrónico** deben ser contestados de inmediato.

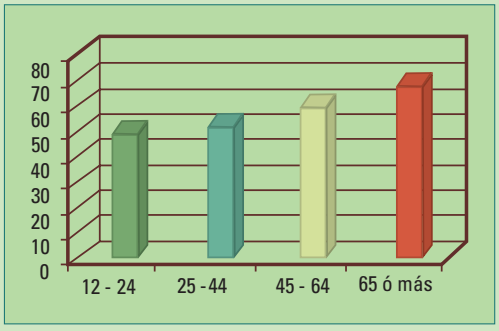


Fig 3.1. *Porcentaje de respuestas a favor de contestar los mensajes del correo electrónico en forma inmediata por grupo de edad. Fuente: USA Today 19 Ago 2008.*

El diagrama de barras muestra los porcentajes de respuestas afirmativas por grupo de edad (12-24, 25-44, 45-54, 65 ó más años). Los porcentajes obtenidos fueron 53%, 54%, 61%, 71% respectivamente. El artículo en el que se muestran estos datos concluye: **"Los mayores responden más rápido el correo electrónico".**

¿Es una conclusión adecuada? ¿Se trata de una respuesta a otra pregunta?

¿Se vincula la conclusión con la pregunta planteada?

Habría que saber un poco más sobre cómo se hizo la encuesta:

1. ¿Cuántas personas respondieron en cada grupo? Si la cantidad de personas de los grupos de edades es muy diferente, los porcentajes (los estadísticos) calculados en cada uno de ellos tiene diferente confiabilidad, y podría no tener sentido compararlos.

2. ¿Cómo fueron elegidas esas personas? Podrían ser las primeras 100 en la puerta de una escuela.

3. ¿Qué características tienen los consultados? Podrían ser todas mujeres.

4. ¿Representan esas personas a la mayoría de la población con su misma edad como para concluir que los mayores responden más rápido el correo electrónico? Si las personas encuestadas fueron las 100 primeras personas que salieron de la escuela es muy posible que sus características no representen a la mayoría de la población.

5. ¿Los usuarios de Internet mayores de 65 años son como el resto de sus contemporáneos o son distintos? Es posible que los usuarios de Internet que tienen más de 65 años tengan diferentes inquietudes que los hombres y mujeres de su misma edad y no sean usuarios de Internet.

6. ¿Opinar que un correo electrónico debe responderse de inmediato, es lo mismo que efectivamente hacerlo? La pregunta de la encuesta se refiere a si los mensajes del correo electrónico **deben ser** contestados de inmediato. Eso no implica que los que contesten por sí, necesariamente lo hagan. Por lo tanto, la conclusión "Los mayores responden más rápido el correo electrónico", **no puede obtenerse a partir de la encuesta realizada.**

□ 3.2 Publicidad

Veremos dos ejemplos que aparecen habitualmente en los medios de comunicación. Se apoyan en resultados estadísticos, sin embargo, sus conclusiones no se ajustan a los mismos.

3.2.1 Crema reductora

La **publicidad de una crema reductora** afirma: ¡3 cm menos!

En letra chica dice:

*"Primeros resultados medibles en muslos, caderas, panza y cintura luego de 4 semanas de uso** 80% de mujeres convencidas**
**Reducción de hasta 3cm en el contorno de muslos, caderas, panza y cintura entre las 4 y 8 semanas de uso*
***Testeado en 168 mujeres durante 3 semanas"*

¿Se encuentra una justificación suficiente a la afirmación ¡3 cm menos!, en letra chica? Veamos.

- ¿Qué significa una reducción de "hasta 3 cm"? Significa que como máximo se obtendrá una reducción de 3 cm, pero puede ser menor. Aunque no ocurriera una reducción, o

incluso si se diera un aumento de la cintura no nos estarían mintiendo.
- Una información más útil podría ser el rango de valores obtenidos. No es lo mismo que las reducciones se encuentren entre 0 y 3 cm, a que estén entre 2,5 y 3 cm. En el primer caso podría haber muchas personas a las que la crema no les hizo absolutamente nada, y en el segundo, la crema parece haber sido efectiva para todos.
- Además, la publicidad sugiere un uso de entre 4 y 8 semanas, para ver si logramos un resultado, cuando la crema fue testeada durante 3 semanas. ¿Cómo pudieron llegar a esa conclusión?

Muchas veces en los medios de difusión, como ocurre en este ejemplo, **se refuerza una afirmación con argumentos estadísticos falsos.**

3.2.2 Pasta dental

La publicidad de una pasta dental afirma que 4 de cada 5 odontólogos recomiendan una marca determinada. ¿Cuántos dentistas fueron encuestados? No se sabe. Porque la publicidad no lo dice. ¿Por qué importa saber la cantidad de respondentes? La fiabilidad del resultado depende de la cantidad de información que se analice, siempre que ésta sea de buena calidad (veremos en los próximos capítulos cómo se produce información de buena calidad).

Cuando los anunciantes dicen "4 de 5 odontólogos" es posible que en realidad hayan sido 5 los odontólogos encuestados, o ninguno si es que inventaron el resultado. También pueden haber sido 5.000 y 4.000 recomendaron dicha marca, que no es lo mismo. No se sabe cuántos dentistas realmente recomiendan esa pasta.

□ 3.3 Razón, tasa y porcentaje

Los estadísticos que se utilicen para describir cantidades pueden hacer una diferencia, respecto a las conclusiones que se obtienen. Primero, veremos algunas definiciones para, finalmente, desarrollar un ejemplo sobre la medición de los accidentes de tránsito.

3.3.1 Definiciones

Una razón es el **cociente entre dos cantidades.** Por ejemplo, "La razón de niñas a niños es de 3 a 2" significa que hay 3 niñas por cada 2 niños. No debe entenderse que sólo hay 3 niñas y 2 niños en el grupo. Las razones se expresan utilizando los términos más bajos para simplificar lo más posible. Así, esta razón expresa la situación de un curso de 25 alumnos con 15 niñas y 10 niños, o de un colegio con 300 chicas y 200 chicos.

Una tasa (o velocidad) es un **cociente que refleja una cierta cantidad por unidad.** Por ejemplo, un automóvil se desplaza a 45 km por hora (la unidad es una hora), o la tasa de robos en un barrio, 3 robos por cada 1.000 hogares (la unidad es 1.000 hogares).

Un porcentaje es un **número entre 0 y 100** que mide la **proporción** de un total. Por ejemplo, cuando decimos que una camisa tiene un 10% de descuento, si el precio original (el total) es $ 90, el descuento es de $ 9. Si decimos que el 35% de la población está a favor de un período de cuatro días de trabajo a la semana, y la población tiene 50.000 habitantes, entonces son 17.500 (50.000 x 0,35 = 17.500) los que están a favor. La proporción de los que están a favor es 0,35.

- Un porcentaje del 35% es lo mismo que una proporción de 0,35
- Para convertir **un porcentaje** en **una proporción,** se **divide** al porcentaje por 100.
- Para convertir una proporción en un porcentaje, se multiplica la proporción por 100.

3.3.2 Variaciones relativas

Cuando un porcentaje se utiliza para determinar un aumento o reducción relativa (relativa al valor inicial), se denomina **variación porcentual.**

Supongamos que la cantidad de accidentes por año en una ciudad pasó de 50 a 60, mientras que la cantidad de accidentes en otra ciudad pasó de 500 a 510. Ambas ciudades tuvieron un **aumento** de 10 accidentes por año, pero para la primera ciudad, esta diferencia como porcentaje del número inicial de accidentes, es mucho mayor.

Variación porcentual: se toma el valor "después de" y se le resta el "antes de", luego se divide ese resultado por el "antes de". Así, se obtiene una proporción. Para transformarla en un porcentaje se multiplica el resultado por 100.

Para la primera ciudad, esto significa que la cantidad de accidentes aumentó en un

$$\frac{60 - 50}{50} = \frac{10}{50}$$

$$= 0{,}20 \text{ ó } 20\%$$

Para la segunda ciudad, este cambio refleja sólo un aumento del 2%, pues

$$\frac{510 - 500}{500} = \frac{10}{500}$$

$$= 0{,}02 \text{ ó } 2\%$$

Si una ciudad pasó de 50 a 40, mientras que en otra la cantidad de accidentes pasó de 500 a 490, ambas ciudades tuvieron una **reducción** de 10 accidentes. Calculemos las variaciones en este caso:

$$\frac{40-50}{50} = \frac{-10}{50} \quad \textbf{y} \quad \frac{490-500}{500} = \frac{-10}{500}$$

$$= -0{,}20 \text{ ó } -20\% \qquad = -0{,}02 \text{ ó } -2\%$$

Las reducciones se reflejan en variaciones porcentuales negativas.

Las variaciones relativas se pueden expresar como variaciones porcentuales o proporciones.

3.3.3 ¿Cantidades o tasas?

El resultado puede ser diferente según que estadístico se elija. Veamos un ejemplo.

3.3.3.1 Accidentes de tránsito

¿Cómo medimos los accidentes de tránsito? Veamos dos maneras de analizar las estadísticas sobre los accidentes de tránsito, mostrando dos aspectos diferentes de la misma historia.

Muchas veces el análisis puede utilizarse con fines políticos. Un candidato puede argumentar que los accidentes fatales se han reducido durante su mandato y su contrincante que han aumentado. A partir de una misma realidad, ¿cómo pueden los dos candidatos decir que la cantidad de accidentes fatales evoluciona en dos direcciones diferentes?

Consideremos los datos de la tabla 3.1, que muestran la cantidad total de víctimas mortales por accidentes de tránsito en el lugar del hecho, en la Argentina desde el año 2.000 hasta el año 2007, de acuerdo con el Registro Nacional de Antecedentes de Tránsito (R.e.N.A.T.).

La cantidad de víctimas mortales se redujo desde el año 2.000 hasta el 2004. A partir de ese año, los accidentes comenzaron a aumentar. Podría decirse que en el 2007 estuvimos peor que en el 2001, con 135 muertes más (4.175 contra 4.040). Pero, ¿**la cantidad de víctimas mortales** es la medida **adecuada** para describir el problema?

Estas cifras no dicen toda la historia. Una parte importante de la información ha quedado fuera. Aumentó la cantidad de accidentes fatales, pero también aumentó la cantidad de vehículos circulantes. Ante iguales condiciones de conducción, es razonable esperar que si aumentan los vehículos circulantes aumentarán los accidentes de tránsito y, por lo tanto, las víctimas fatales. Para poner el problema en perspectiva es necesario incluir en el análisis, tanto la cantidad de vehículos circulantes como la cantidad de muertes. ¿Cómo se hace?

El Registro Nacional de Antecedentes de Tránsito (R.e.N.A.T.) publica además de la cantidad de muertes, la tasa de muertes por cada 100.000 vehículos en circulación.

VÍCTIMAS MORTALES POR ACCIDENTES DE TRÁNSITO EN EL LUGAR DEL HECHO EN LA REPÚBLICA ARGENTINA (2000 - 2007) TABLA 3.1

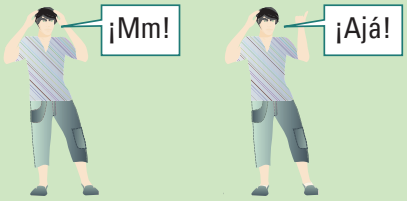
Año	Cantidad total muertes	Cantidad de vehículos circulantes	Tasa c/100.000
2000	4.316	6.799.114	63,48
2001	**4.040**	6.937.355	**58,24**
2002	3.830	7.005.406	54,67
2003	3.690	7.102.855	51,95
2004	3.047	7.355.731	41,42
2005	3.378	7.717.513	43,77
2006	3.842	7.923.726	48,49
2007	**4.175**	**7.995.043**	**52,21**

Fuente: *http://www.renat.gov.ar/Estadistica.htm*

Comparando las tasas de muertes, nuevamente entre los años 2007 y 2001 vemos que se redujo (52,21 contra 58,24).

Una tasa es un cociente. Refleja una cantidad dividida por una cierta unidad. Por ejemplo una velocidad, espacio / tiempo (km/hora), es una tasa.

¿Cómo dijo?



¿Qué significa una tasa de 52,21 **muertes** por accidentes **cada 100.000 vehículos?** 52,21 es la cantidad de muertes y la unidad en este caso es 100.000 vehículos.

¿Cómo se obtiene?

$$\frac{\text{cantidad de muertes}}{\text{cantidad de vehículos}} \times 100.000 = \frac{4.175}{7.995.043} \times 100.000$$

$$= 52,21$$

Tenemos que multiplicar por 100.000 porque $\frac{\text{cantidad de muertes}}{\text{cantidad de vehículos}}$ es la cantidad de accidentes que le corresponderían a un vehículo.

¿Cantidades o tasas?

Dependiendo del estadístico utilizado para resumir la información, para este caso cantidades o tasas, se pueden obtener conclusiones opuestas, como las que obtuvimos al comparar los años 2001 y 2007 en relación a los accidentes fatales.

¿Cuál es el estadístico correcto? Depende.

Muchas veces la respuesta en un ámbito **estadístico** es: depende.

Depende de la pregunta que queramos responder. Si nos interesa evaluar el éxito de la política de educación vial deberíamos utilizar tasas de muertes; pero si organizamos el servicio de ambulancias importa la cantidad de accidentes y no los motivos (aumento de vehículos, aumento de la población, aumento de la cantidad de conductores imprudentes).

En el 2007 tuvimos **4.175** víctimas mortales por accidentes de tránsito, esta cantidad de muertes equivale a la caída de un avión jet sin sobrevivientes cada quince días (un avión jet lleva aproximadamente 150 pasajeros). ¡Eso es mucho!

"Aquí estamos utilizando el término estadístico con dos significados diferentes:

1. procedimiento para obtener un número
2. disciplina"

3.4 Actividades y ejercicios

1. El 7 de septiembre de 2008 podía leerse en un diario: "El producto bruto de Brasil es 1.313 billones de dólares. Es cuatro veces superior al PBI que hoy tiene la Argentina". ¿Le parece adecuada esa conclusión teniendo en cuenta que la población de Brasil era para esa fecha 189,3 millones de habitantes, mientras que la nuestra era de 39,7 millones?

2. Un investigador señala que la cantidad de accidentes en su ciudad es mayor entre las 18 h y 20 h (tarde tarde) que entre las 14 h y las 16 h (tarde temprana). Concluye que la fatiga juega un papel muy importante en los accidentes de tránsito, porque los conductores están más cansados durante la tarde tarde que durante la tarde temprano. ¿Considera que esta conclusión está bien justificada?

3. Halle 3 ó más **noticias** o artículos de opinión que presenten, tasas, proporciones o porcentajes (o algún otro cálculo de tipo estadístico) para justificar un punto de vista.

4. Halle 3 ó más avisos publicitarios que muestren resultados de estudios estadísticos para resaltar la efectividad o preferencia de un producto.

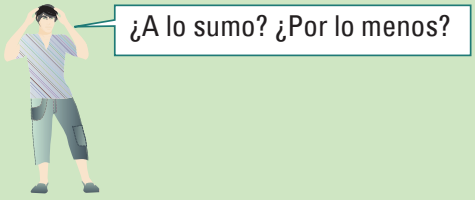
5. Realice las preguntas que considere necesarias para evaluar las siguientes afirmaciones de un aviso publicitario anunciando un producto contra la celulitis:

- En 15 días piel de naranja menos visible*
- Piel más lisa 86%**
- -1,9 cm en 4 semanas*

**Test clínico en 50 mujeres. **Autoevaluación sobre 44 individuos.*

6. Explique las siguientes frases:

- Le puedo pagar a lo sumo $ 500 por ese trabajo.
- Le voy a pagar como mínimo $ 500 por ese trabajo.
- Quiero que vuelvas como máximo a las 11 de la noche.
- Se presentaron por lo menos 10 personas para el puesto de encargado de control de calidad.
- No más de 10 personas se presentaron para el puesto de chofer.



7. “Un chico de 8 a 12 años puede perder hasta un litro de transpiración durante dos horas de actividad un día caluroso”, afirma una publicidad. Nos preguntamos:

- ¿cómo se podrá llegar a esa conclusión?
- ¿cómo será para los de 13 a 16 años?
- Aunque sea complicado, proponga algún procedimiento para estimar cuanto líquido puede perder un chico por transpiración durante dos horas.
- “Hasta dos litros” ¿significa que puede:

    - no perder nada?
    - perder 3 litros?
    - perder 2 litros?
    - perder 1 litro?
    - perder 1,5 litros?
    - perder 2,5 litros?

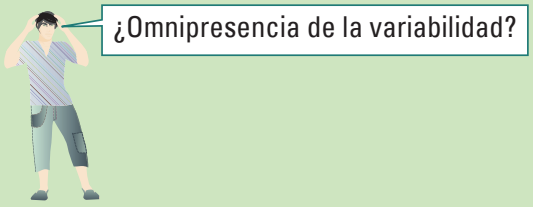
4. Herramienta para la ciencia

La estadística interviene activamente en todas las etapas que componen el método científico.

Aunque para el método científico no exista una secuencia única, señalamos los siguientes pasos generales:

- planteo de preguntas,
- planificación y realización de los estudios,
- recolección de datos,
- análisis de la información,
- obtención de las conclusiones.

Aunque en sí misma es una ciencia que se dedica al desarrollo de nuevos métodos y modelos, la estadística es una disciplina que ofrece un conjunto de ideas y herramientas para el tratamiento de datos. En este sentido, podemos decir que se trata de una disciplina metodológica, su necesidad se deriva de la omnipresencia de la variabilidad.



Los estadísticos trabajan junto a expertos en diferentes disciplinas.

La estadística está involucrada en el proceso que va desde la recolección de la evidencia, el planteo de las preguntas, hasta hallar las respuestas.

Lo importante es que haya preguntas. Para hallar respuestas se pueden seguir diferentes caminos.

Toda investigación comienza con preguntas como:

- ¿Hace mal comer papas fritas?
- ¿Cuánto cuesta enviar un/a niño/a al colegio?

- ¿Quién ganará las próximas elecciones de un club de fútbol?
- ¿Está el peso del cerebro, relacionado con la inteligencia?
- ¿Es posible tomar demasiada agua?

Aunque ninguna de las preguntas anteriores se refiere directamente a números, para responderlas se requiere del uso de datos y de un procedimiento estadístico.

Veamos un ejemplo:

Supongamos que una investigadora quiere determinar quién ganará las próximas elecciones para presidente del Club Grande de Fútbol (58.210 socios), y supongamos también que ya tiene la pregunta, para responderla deberá seguir los siguientes pasos:

- **Determinar el grupo de personas a participar del estudio**

  Puede utilizar la lista de socios en condiciones de votar.

- **Recolectar los datos**

  Este paso es más difícil. No se puede preguntar a todos los socios si van a ir a votar, y en el caso afirmativo, ¿a quién? Supongamos que alguien dice que irá a votar y declara a quién votará: ¿realmente irá esa persona a votar el día de las elecciones?; ¿dirá esa persona a quién piensa votar realmente? ¿Y si el día de la votación cambia de opinión?

- **Organizar, resumir y analizar los datos**

  Luego de recolectar los datos, la investigadora necesita organizar, resumir y analizarlos. Habitualmente, esta parte de su trabajo se reconoce como una tarea para estadísticos.

- **Obtener los resúmenes, tablas y gráficos, realizar el análisis, conclusiones tratando de responder la pregunta del investigador.**

  Presentaremos los tipos de resúmenes, tablas y gráficos que podrá utilizar a partir del capítulo 7. Aunque no podamos todavía describir detalladamente el análisis, sabemos que la investigadora no podrá tener un 100% de confianza en que sus resultados sean correctos, porque no le ha podido preguntar a todas las personas y, además algunas pueden cambiar de opinión. Pero sí es posible tener una confianza cercana al 100%, digamos 95% de que la estimación es correcta. De hecho, si ha tomado una muestra representativa (sección 5.4.1) de –por ejemplo- unas 600 personas, de manera que todos los socios tienen igual chance de ser elegidos (muestra insesgada), puede tenerse un resultado preciso con un margen de error de más o menos 4%. Siempre existe la posibilidad de que la conclusión de un estudio sea errónea. Veremos más adelante (sección 10.1) que **el margen de error sólo depende del tamaño de la muestra** y no del tamaño de la población. También veremos el significado del porcentaje de confianza y cuán preciso se espera que sea un resultado.

- **Nuevas preguntas**

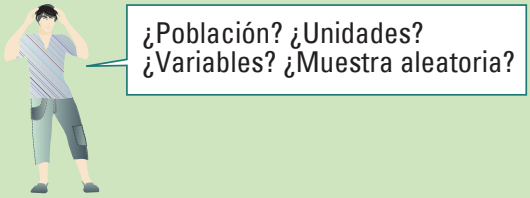
  Cuando se concluye una investigación y se han contestado las preguntas, los resultados suelen llevar a nuevas preguntas. Podría averiguarse porqué los socios jóvenes prefieren al candidato Rolando Forzudo y los socios mayores a su oponente. A Forzudo podría interesarle estudiar qué factores hacen que los jóvenes realmente vayan a votar.

Hemos dicho que la necesidad de la estadística se deriva de la omnipresencia de la variabilidad. Pero, ¿dónde se encuentra la variabilidad en el ejemplo de la encuesta sobre la preferencia del candidato? Hay varias fuentes (o motivos) que producen variabilidad. La primera, y la más importante, es que no todos los socios piensan igual, si lo hicieran alcanzaría con saber que piensa uno de ellos. La segunda resulta porque las personas cambian de opinión, si esta fuente de variabilidad es muy grande, la validez del resultado depende de cuán cerca de las elecciones se realice la encuesta. La tercera se debe a que los encuestados pueden mentir (sección 6.3.2.).

5. Vocabulario - jerga

Los conceptos requieren de palabras específicas para ser identificados.

La estadística tiene su propio vocabulario. Veremos algunos términos básicos, que volveremos a encontrar más adelante, además, seguiremos incorporando términos a lo largo de todo el libro.



Con la intención de fijar ideas, retomemos la investigación para saber quién ganará las próximas elecciones como presidente del Club Grande de Fútbol.

El primer paso es determinar el **grupo de personas a ser estudiadas,** o sea determinar la **población en estudio.** En este caso es la totalidad de los socios del Club Grande de Fútbol en condiciones de votar.

El segundo paso es **recolectar los datos.** Aquí aparecen varias cuestiones que nos permiten ilustrar más términos específicos. ¿Cuáles **individuos** serán encuestados?, esto es, ¿cuál será la **muestra**? ¿Se los elegirá en forma **aleatoria** de manera que todos los socios tengan la misma oportunidad de ser seleccionados? ¿Qué **variables** (edad, género) serán importantes en relación al tema central de la encuesta (candidato preferido)?

□ 5.1 Unidades muestrales

A los objetos de interés de un estudio se los denomina **unidades muestrales** o simplemente unidades. Muchas veces, las unidades muestrales son **individuos**: tornillos, personas, tubos de pasta dentífrica, lamparitas. Otras veces, las unidades están compuestas por **muchos individuos:** ciudades, escuelas, lotes (de tornillos) etc.

□ 5.2 Variables

Las **variables** son **características** que pueden **cambiar** de una **unidad** muestral a otra, como la **edad** de las personas, la población de cada ciudad, el **porcentaje** de alumnos reprobados de una escuela, la **preferencia** de una comida balanceada para un animal, la **intensidad** de emisión de rayos X de cada televisor, la **capacidad** de almacenamiento de un disco rígido, la **longitud** de un tornillo, la **duración** o el consumo de una lamparita.

> ! No confundir una **unidad** muestral como **objeto** completo y diferenciado que se encuentra dentro de un conjunto (una docena tiene doce unidades) con las unidades que se utilizan para valorar una magnitud (el metro es una unidad de longitud).

□ 5.3 Población

Para cualquier pregunta que interese responder, primero es necesario dirigir la atención a un **grupo particular de unidades** muestrales: personas, ciudades, animales, televisores, discos rígidos, tornillos o lamparitas.

- ¿Qué piensan los porteños sobre el Sistema de Evaluación Permanente de Conductores?
- ¿Qué porcentaje de familias de la ciudad de Santa Fe tienen mascotas?
- ¿Cuál es la expectativa de vida de los diabéticos?
- ¿Qué porcentaje de todos los tubos de pasta dentífrica son llenados de acuerdo a sus especificaciones?
- ¿Cuál es la duración promedio de las lámparas de bajo consumo de una determinada marca?
- ¿Los jóvenes deportistas consumen menos alcohol que los sedentarios?

En cada uno de los ejemplos, se plantea una pregunta y se puede identificar uno o más grupos específicos de unidades que interesa estudiar: los porteños (habitantes de la ciudad de Buenos Aires), las familias de la ciudad de Santa Fe, los diabéticos, los tubos de pasta dentífrica, las lámparas de bajo consumo, los deportistas y los sedentarios.

Se llama **población a todo el grupo de unidades muestrales** (generalmente son individuos) que interesa estudiar con el fin de responder una pregunta de investigación. Las poblaciones, sin embargo, pueden ser difíciles de definir. En un buen estudio, los investigadores deben **definir la población con toda claridad.**

La pregunta respecto a si los jóvenes que practican deportes consumen menos alcohol, sirve de ejemplo para ver lo difícil que puede ser definir con precisión la población. ¿Cómo definir un joven? ¿Los menores de 18 años de edad? ¿Los menores de 30 años? ¿Cómo definiría un sedentario? ¿Interesa estudiar los jóvenes de la República Argentina

o los de todo el mundo? Los resultados pueden ser diferentes para los menores de 18 que para los mayores, para los latinoamericanos comparados con los europeos, y así otras clasificaciones.

Muchas veces, los investigadores quieren estudiar y sacar conclusiones sobre una **población amplia** pero, con el fin de ahorrar tiempo, dinero, o simplemente porque no se les ocurre nada mejor, sólo estudian una **población muy restringida.** Esto puede conducir a serios problemas al momento de sacar conclusiones.

Supongamos que un profesor universitario quiere estudiar si los jóvenes que practican deportes consumen menos alcohol. Basa su estudio en un grupo de sus alumnos, que participan porque al hacerlo se les da cinco puntos adicionales en su puntaje final. Este grupo de alumnos constituye una muestra; pero los resultados no pueden generalizarse a toda la población de jóvenes, ni siquiera a todos los estudiantes.

□ 5.4 Muestra

¿Qué hacemos para probar la sopa? Revolvemos la olla con una cuchara, sacamos una porción -una muestra- la saboreamos y sacamos una conclusión sobre toda la sopa de la olla sin haber en realidad probado toda. Si la muestra ha sido tomada adecuadamente -sin elegir tramposamente la parte buena- tendremos una buena idea del sabor de la totalidad de la sopa. Esto se hace en estadística, más específicamente en **inferencia estadística.**

Los investigadores quieren averiguar algo sobre una población, pero no tienen tiempo o dinero para estudiar a todos los individuos que la conforman. Por lo tanto, ¿qué hacen? Seleccionan una **cantidad pequeña de unidades muestrales de la población** (esto se llama una **muestra**), estudian esas unidades, generalmente individuos, y utilizan esa información para sacar conclusiones sobre toda de la población.

5.4.1 Muestra representativa

Nos interesa obtener **"buenas muestras".**

Una buena **muestra** debe ser **representativa** de la población. Esto significa, que todas las características importantes de la población tienen que estar en la muestra en **la misma proporción que en la población.**

Una muestra tiene, en pequeño y lo más parecidas posibles, las características de la población.

Podremos sacar conclusiones respecto de la población total a partir de una muestra -esto es, realizar una inferencia-, para todas aquellas características en las cuales la muestra representa a la población.

Ejemplo:

Consideremos un ejemplo simple, una población constituida por personas que difieren entre sí en una única característica con dos categorías:

Característica: el peso
Categorías: gordo, flaco

La figura 5.1 muestra una población hipotética que tiene 18 individuos que son gordos, o flacos (no hay gente con peso normal en esta población) y una muestra representativa.

En la figura 5,1 podemos ver la población total y la muestra. Respecto de la población, podemos decir que: 5 de cada 9 personas son gordas. Esta relación se repite en la muestra representativa.

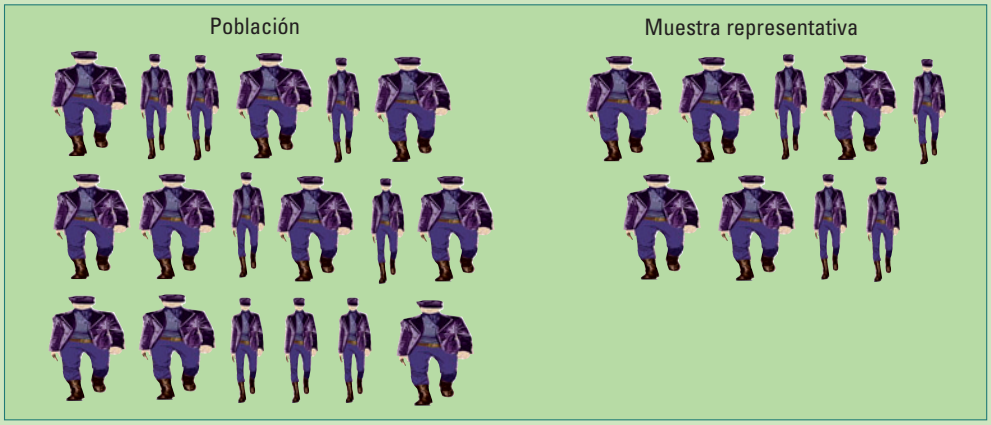


Figura 5.1 Muestra representativa de una población que sólo tiene una característica, el peso, con dos posibilidades: gordo y flaco.

En la vida real, es muy difícil que una muestra tenga proporciones idénticas a las poblacionales, pero **deberían ser muy parecidas** en todas las características que se puedan conocer.

Si se quiere realizar un estudio para averiguar a qué edad caminan los bebés de la Argentina, la muestra debería tener una distribución geográfica, por provincia o región, similar a la del último censo disponible de población. Si se considera que hay otros factores que pueden influir además de la región, por ejemplo el tipo de vivienda (casa o departamento), la muestra también tendrá que ser representativa de esos otros factores.

6. Muestreo

La forma en que se realiza la selección puede hacer la diferencia. Es más fácil obtener muestras malas que buenas.

No todo es tan simple como tomar sopa.

En la Sección 5.3 consideramos un estudio, realizado por un profesor universitario entre sus alumnos, para evaluar si los jóvenes que practican deportes consumen menos alcohol. Este es un ejemplo de participación voluntaria en un estudio, la muestra no es representativa de la población de interés.

Recordemos un ejemplo de la Sección 3.1. Interesaba conocer las opiniones respecto a si el correo electrónico debe responderse lo más rápido posible o no. Si la encuesta fue realizada vía el correo electrónico, las opiniones representan únicamente a los que tienen correo electrónico y les interesó responder la encuesta.

La próxima vez que se encuentre con un resultado de un estudio, averigüe qué composición tenía la muestra y pregúntese si la muestra representa a la población que interesa o a un subgrupo más restringido.

□ 6.1 Muestreo aleatorio simple

Es bueno que la **muestra** se seleccione en forma **aleatoria;** esto significa que:

Cada uno de los individuos de la población tiene la misma oportunidad de ser seleccionado.

- Se utiliza algún mecanismo probabilístico para elegirlos.
- La gente no se selecciona a sí misma para participar.
- Nadie en la población es favorecido en el proceso de selección.

Muestra aleatoria simple: Una muestra aleatoria simple es la que se obtiene a partir de un mecanismo que le da a cada una de las unidades muestrales la misma probabilidad de ser elegida.

El muestreo aleatorio (el proceso por el cual se obtiene una muestra aleatoria) comienza con una lista de **unidades muestrales** de la que se extraerá la muestra. Esta lista se llama **marco muestral.** Idealmente, el marco muestral debería contener la lista de la totalidad de las unidades muestrales.

El **muestreo aleatorio simple** tiene dos propiedades que lo convierten en el procedimiento por excelencia de obtención de muestras.

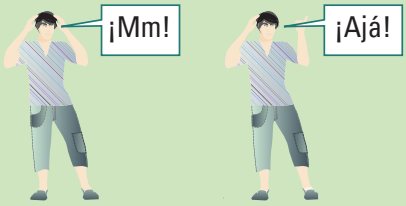
- Todas las unidades tienen la misma oportunidad de ser elegidas (es insesgado).
- La elección de una unidad no influye sobre la elección de otra (independencia).

El Instituto Nacional de Estadísticas y Censos - INDEC - realiza periódicamente **censos** para registrar las características básicas sobre población y vivienda, actividad económica y agropecuaria de nuestro país. Las unidades relevadas en los censos proveen el **marco muestral** para las **encuestas** que realiza durante los períodos intercensales.

Se espera que el muestreo aleatorio provea muestras representativas de la población.

Mediante un censo se intenta registrar todas las unidades muestrales de la población para proveer el marco muestral. Si se trata de un censo de población, deberán localizarse todas las personas. Si se trata de un censo económico, se registrarán todos los locales comerciales y productivos. Una vez que se dispone del marco muestral se abre la oportunidad de seleccionar la muestra.

Por otra parte, es necesario aclarar que una unidad muestral puede contener muchos individuos. Una escuela, con sus alumnos, puede ser una unidad muestral. El objetivo del estudio pueden ser las escuelas (por ej. interesa conocer la superficie cubierta por alumno) o ser los alumnos (por ej. interesa conocer el rendimiento en educación física).



¿Cómo? ¿Una unidad muestral puede estar constituida por muchos individuos?

Volvamos al ejemplo de la encuesta sobre la preferencia del candidato a presidente del Club Grande de Fútbol. Utilicemos la lista actualizada de todos los socios como marco muestral con los números de socio para identificarlos. Si se decide que 1 de cada 6 socios entrarán en la muestra podemos arrojar un dado tantas veces como socios tenemos en la lista y si sale 1 el socio es seleccionado.

TABLA 6.1

Socio Número	Número aleatorio	Socio Número	Número aleatorio	Socio Número	Número aleatorio	Socio Número	Número aleatorio
1495	4	**1.501**	**1**	1.507	4	1.513	4
1496	8	1.502	6	1.508	4	1.514	7
1497	8	1.503	3	1.509	3	1.515	8
1498	7	1.504	7	1.510	8	1.516	8
1499	9	**1.505**	**1**	**1.511**	**1**	**1.517**	**1**
1500	5	1.506	7	1.512	7	1.518	3

Con este procedimiento, seleccionamos los socios no: 1.501, 1.505, 1.511 y 1.517 mediante un **muestreo aleatorio simple.**

También podríamos utilizar un programa de computadora para generar números entre 1 y 6 en forma aleatoria, sin necesidad de arrojar un dado.

Muestra aleatoria simple en dos pasos :
Paso 1. Se asigna una etiqueta numérica a cada individuo de la población.
Paso 2: Se utilizan números aleatorios para seleccionar las etiquetas al azar.

En la práctica, el primer paso del procedimiento es el más difícil. Esta dificultad da lugar a **muestreos alternativos** que **no** son **válidos** desde el punto de vista del análisis estadístico. Veremos algunos en la próxima sección.

□ 6.2 Muestras malas

Todos los días encontramos ejemplos de **muestras malas:**

- Cuando se pide a los oyentes de un programa de radio que voten por tal o cual cantante, llamando por teléfono o enviando un mensaje de correo electrónico, se trata de **muestras de respuesta voluntaria.** Las encuestas de opinión en las que se llama, o se escribe, por propia iniciativa son ejemplos de muestras de respuesta voluntaria, poco satisfactorias desde un punto de vista estadístico.

Cada vez que mire los resultados de un estudio, busque la frase "muestra aleatoria". Si la encuentra, hile más fino para averiguar cómo fue obtenida y si en realidad fue elegida en forma aleatoria.

- Otro tipo de muestra mala es la **muestra de conveniencia.** Si una pedagoga elige a sus propios alumnos, del último año de la escuela secundaria en la que trabaja, para evaluar un cambio en el método de enseñanza, los resultados no se podrán extender más allá de ese grupo.

□ 6.3 Sesgo

Alguna vez escuchamos **el sesgo es malo.** Pero, ¿qué es el sesgo? Es un favoritismo de alguna etapa del proceso de recolección de datos **beneficiando algunos resultados, perjudicando otros y desviando las conclusiones en direcciones equivocadas.**

> Cuando alguna etapa del proceso de recolección de datos está sesgada, **utilizar una muestra grande no corrige el error,** simplemente lo repite.

Los datos en un estudio pueden estar sesgados por muchos motivos. A continuación, veremos algunos de ellos.

6.3.1 Sesgo por elección de la muestra

6.3.1.1 Muestras por conveniencia

Exprimir las naranjas que se encuentran a la vista, en la parte de arriba del cajón, es un ejemplo de muestra de conveniencia. Las entrevistas en los centros comerciales (shopping) son otro ejemplo, porque los fabricantes y las agencias de publicidad suelen recolectar información respecto a los hábitos de compras de la población y el efecto de sus publicidades en grandes centros de compras. Obtener una muestra de esta manera es rápido y económico, pero la gente que contactan no es representativa de la mayoría de la población.

6.3.1.2 Muestras con sesgo personal

Por simpatía, gusto o interés, quien está realizando la encuesta puede preferir encuestar a cierto tipo de personas y no a otras. Por ejemplo, es posible que un encuestador joven tienda a buscar chicas bonitas para preguntarles.

6.3.1.3 Muestras de respuesta voluntaria

Surgen a partir de los individuos que se ofrecen voluntariamente a participar. Se trata, por ejemplo, de las que alimentan las votaciones organizadas por programas de radio, televisión o de algún sitio de Internet. No producen resultados que tengan algún significado en relación a la opinión de la población en general. Los participantes voluntarios, que por algún motivo decidieron participar, suelen tener opiniones más polarizadas.

6.3.2 Sesgo de respuesta

6.3.2.1 Debido a la presentación de las preguntas

Las diferentes palabras con las que se puede presentar una misma pregunta suele ser una fuente importante de sesgo en las respuestas.

En un curso de manejo organizado por un automóvil club se proyectó una película sobre un accidente de tránsito a dos grupos de alumnos. Ambos grupos eran similares respecto de la edad y el género. Al finalizar la proyección se preguntó:

- Al primer grupo: ¿a qué velocidad piensa que los dos autos chocaron? El promedio de las respuestas fue de 50,9 km/h.
- Al segundo grupo: ¿a qué velocidad piensa que los dos autos se colisionaron? El promedio de las respuestas fue de 65,9 km/h.

Ambos grupos vieron la misma película. El uso de la palabra **colisionaron** aumentó las estimaciones de la velocidad del accidente en **15 km/h,** esto es un aumento del 29,5 %

El sesgo debido a la forma en que se presenta una pregunta puede ser **intencional** o **no intencional.**

Las preguntas "¿No está usted harto de pagar impuestos para que todo siga igual de mal?" y "¿Le parece importante que se paguen impuestos para mejorar la educación, los servicios de salud y la seguridad?", que apuntan al pago de impuestos, seguramente tendrán resultados muy diferentes. Ambas preguntas conllevan un **sesgo intencional.**

Una encuesta dirigida a alumnos de 7mo. grado que pregunte: "¿Cuáles son las 5 personas grandes que le gustarían conocer personalmente?" tendrá diferentes lecturas. Algunos de los alumnos podrán interpretar que se trata de personas mayores de edad, otros que son altos, otros que se refiere a gordos o tal vez a grandes estrellas de cine, de rock, políticos o deportistas, generando un sesgo **no intencional.**

6.3.2.2 Para tratar de agradar

A la gente no le gusta mostrarse con ideas que no están bien vistas socialmente. Por ejemplo, cuando esté cara a cara con un encuestador o llenando un formulario no anónimo, un varón evitará una respuesta que parezca machista, o una mujer responderá tratando de ocultar algún prejuicio.

6.3.2.3 Por recuerdo

Si la pregunta está referida a un acontecimiento ocurrido algún tiempo atrás, la respuesta tendrá un sesgo por recuerdo. Por ejemplo, si se le pregunta a una madre a qué edad comenzaron a caminar sus hijos, la veracidad y precisión de la respuesta dependerá de las características personales de la madre.

6.3.2.4 Por no respuesta

Algunas veces las personas que han sido seleccionadas para una encuesta son muy difíciles de localizar o simplemente se niegan a responder. **Los individuos que no responden pueden ser muy diferentes de los que sí lo hacen.** Este tipo de sesgo se puede reducir sustituyendo a los se niegan a responder por otros individuos con las mismas características de los que no responden, pero suele ser difícil.

6.3.2.5 Por subcubrimiento

Una encuesta telefónica ignora a todos los sujetos que no tienen teléfono. Una encuesta que realiza las entrevistas en hogares ignora a los que viven en la calle.

Cuando mire los resultados de una encuesta que le interesa especialmente, antes de sacar sus propias conclusiones averigüe qué se preguntó, cómo fueron redactadas las preguntas, si las respuestas fueron dadas en forma anónima o no y cuántos se negaron a responder.

Es más fácil obtener muestras malas que buenas.

□ 6.4 Otros tipos de muestreos

6.4.1 Muestreo sistemático

Veamos un ejemplo de la utilidad de este método. Si nos interesa la opinión de las alumnas de una escuela respecto del aumento de las horas destinadas a la práctica de deportes, podríamos entrevistar a las alumnas a la salida y elegir una de cada diez (suponiendo que salgan de a una) hasta que hayan salido todas. De esta manera, si la escuela tiene 227 alumnas, la muestra tendrá 22 alumnas.

Muestreo sistemático: El muestreo comienza con una unidad elegida al azar y a partir de allí continúa cada k unidades. Si n es el tamaño muestral y N es el tamaño de la población entonces k es aproximadamente N/n.

Este tipo de muestreo permite evitar el sesgo personal y es más sencillo que el muestreo aleatorio. Es útil cuando la población está ordenada naturalmente (si no lo está, para utilizar este tipo de muestreo es necesario ordenarla, pero al ordenarla, se pierden las ventajas que tiene).

Por su simplicidad, se suele utilizar para **control de calidad** durante, o al finalizar, la fabricación de diversos productos.

En una producción continua de tubos de pasta dentífrica, se elige un tubo por hora y se lo analiza para verificar que cumple con las especificaciones.

> **Advertencia:** Este muestreo no es adecuado cuando el período de la selección está relacionado con alguna característica que nos interesa evaluar.

Podría ocurrir que cada hora (una hora es el período de la selección) se produzca una leve caída de tensión que hace que los tubos de pasta dentífrica se llenen más o menos. No detectaríamos esa variación con el muestreo cada hora.

Al realizar un muestreo sistemático es importante estar alerta para identificar los factores que puedan estar invalidando los resultados.

6.4.2 Muestreo aleatorio estratificado

En un muestreo estratificado la población se divide en **grupos homogéneos** llamados estratos. Luego se realiza un muestreo aleatorio simple de unidades muestrales dentro de cada estrato.

Los estratos se eligen de acuerdo con los valores conocidos de algunas variables, de manera que haya **poca variabilidad dentro del estrato** (los valores de dichas variables para las unidades de un estrato particular difieren poco), pero que haya **mucha variabilidad entre estratos** (los valores de dichas variables para las unidades de distintos estratos difieren mucho).

Ejemplo 1:

La población de una ciudad podría estratificarse por

- **grupo de edad:** menos de 6 años, entre 6 y 12 años, entre 13 y 18 años y mayores de 18 años.
- **género:** femenino, masculino.

Así obtenemos 8 estratos, dentro de los cuales los individuos tienen 2 características similares: grupo de edad y género. Podríamos realizar un muestreo proporcional a la cantidad de individuos que tiene cada estrato, de manera que el tamaño de la muestra dentro de cada estrato dependa de la proporción de la población total que dicho estrato representa.

Ejemplo 2:

En una encuesta diseñada para conocer la situación de la industria en una provincia podrían utilizarse estratos por tamaño y actividad. Para cada actividad industrial podrían incluirse **todos** los locales industriales con 500 ó más obreros ocupados (**inclusión forzosa** - la muestra los contiene a todos), **la mitad** de los que tuvieran entre 499 y 200, **la cuarta parte** entre 199 a 50 y **1 de cada 20** para los de menos de 50. Tendríamos así 4 estratos:

- Estrato 1: Locales con 500 ó más obreros
- Estrato 2: Locales con 499-200 obreros
- Estrato 3: Locales con 199-50 obreros
- Estrato 4: Locales con 50-0 obreros

Si además se dividiera la actividad industrial en dos: 1) industria alimenticia, 2) industria no alimenticia, ¿cuántos estratos tendría la muestra? Tendría 8 estratos, dos por cada uno de los 4 estratos anteriores.

Tres pasos de un muestreo aleatorio estratificado:

- **Paso 1:** las unidades se agrupan en estratos. Los estratos se eligen teniendo en cuenta que estos grupos tienen un interés especial dentro de la población, o porque los individuos en el estrato se parecen mucho.
- **Paso 2:** se establece la proporción de unidades, o **fracción de muestreo,** que se incluirá para cada estrato
- **Paso 3:** dentro de cada estrato se realiza un muestreo aleatorio simple y la **proporción de individuos** que se incluye en la muestra es la establecida en el paso 2. La unión de las muestras de cada estrato constituye la muestra completa.

6.4.3 Muestreo por conglomerados

En un muestro por conglomerados la población se divide en **grupos heterogéneos** llamados **conglomerados**. Luego se realiza un muestreo aleatorio simple en el que las unidades muestrales son los conglomerados.

La idea del agrupamiento para un **muestreo aleatorio por conglomerados** (también llamados aglomerados) es opuesta a la del muestreo estratificado. Interesa que los individuos que componen cada grupo sean lo más heterogéneos posibles y se espera que cada conglomerado sea representativo de la población. Los **conglomerados** son las **unidades del muestreo,** pero las unidades de interés son los individuos dentro de los conglomerados. Se selecciona una muestra aleatoria de conglomerados, y **se observan todos los individuos dentro de cada conglomerado** ó se selecciona una muestra aleatoria simple dentro del conglomerado. Este tipo de muestreo puede tener mejor rendimiento costo-efectividad que un muestreo aleatorio simple, en especial si los costos de traslado son altos.

Ejemplo 1:

Una encuesta de viviendas. Se divide la ciudad en manzanas, se seleccionan las manzanas mediante un muestreo aleatorio simple y se visitan todas las casas de cada manzana seleccionada.

Ejemplo 2:

En un estudio interesa evaluar la capacidad de lectoescritura de alumnos de 7mo grado. Se seleccionarán al azar las escuelas y luego se realizará la prueba en todos los alumnos de 7mo. grado de las escuelas seleccionadas.

Tres pasos de un muestreo aleatorio por conglomerados:

- **Paso 1:** Los individuos se agrupan en conglomerados. Los conglomerados generalmente tienen una proximidad física, pero dentro de cada conglomerado las unidades son heterogéneas.
- **Paso 2:** Los conglomerados son las unidades muestrales. Se establece la proporción de unidades que se incluirá.
- **Paso 3:** Se realiza un muestreo aleatorio simple de conglomerados y se estudian todos los individuos de cada conglomerado seleccionado. El tamaño final de la muestra es la cantidad de individuos que componen todos los conglomerados seleccionados.

6.4.4 Muestreo multietápico

Un muestreo multietápico tiene dos o más pasos y, en cada uno de ellos se aplica cualquiera de los procedimientos de selección anteriores.

Ejemplo 1:

Una encuesta de viviendas. En **la primera etapa** se divide la ciudad en barrios, se toma una muestra aleatoria simple de barrios. En **la segunda etapa,** cada barrio seleccionado en la primera etapa se divide en manzanas, se seleccionan las manzanas mediante un muestreo aleatorio simple, y se visitan todas las casas de cada manzana seleccionada.

Ejemplo 2:

Estudio para evaluar la capacidad de lectoescritura de alumnos de 7mo. grado. En **la primera etapa** se seleccionan al azar las escuelas, y en **la segunda etapa** se selecciona dentro de cada escuela un cierto número de cursos de 7mo. grado. La prueba se realiza en todos los alumnos de 7mo. grado de los cursos seleccionados en la segunda etapa.

6.5 Actividades y ejercicios

1. ¿Cuál es la Población? ¿Cuál es la muestra?
   Para cada uno de los siguientes estudios indicar la población lo más detalladamente posible, es decir describir a los individuos que la componen. Si la información es insuficiente, completarla de la forma que se considere más adecuada. También indicar cuál es la muestra.

   - Una encuesta de opinión contacta a 1.243 adultos y les pregunta, ¿ha comprado un billete de lotería en los últimos 12 meses?
   - Durante la reunión anual del colegio de abogados, todos los presentes (2.500), llenaron una encuesta referida al tipo de seguro que prefería para su automóvil.
   - En 1968 se realizó en Holanda un test de inteligencia a todos los varones de 18 años que estaban realizando el Servicio Militar Obligatorio.
   - El INDEC lleva a cabo la Encuesta Permanente de Hogares (EPH) en la que se encuestan 25.000 hogares para captar información sobre la realidad económico-social de la República Argentina.

2. Voto secreto y obligatorio.

   - ¿Qué tipos de sesgos se pueden producir cuando una elección para presidente se realiza en forma voluntaria?
   - ¿Qué tipos de sesgos se pueden producir si el voto en la Comisión Directiva de un club o en la Cámara de Diputados no es secreto?

3. Se quiere realizar una encuesta entre los alumnos de una escuela secundaria, de 2.500 alumnos (500 alumnos por cada año, de 1ro. a 5to.), utilizando una muestra de tamaño 100. El propósito de la encuesta es determinar si a los/as alumno/as les interesa discutir el siguiente tema: "Debe reducirse la edad de imputabilidad penal para los menores de edad, que establece la ley nacional 22.278, a dieciséis años de edad; como respuesta al incremento en la cantidad de delitos graves cometidos por jóvenes y adolescentes".

4. Indicar cuál es el tipo de muestreo realizado en cada caso.

   - Cada alumno escribe su nombre en un papel, lo pone en una bolsa y el director elige 100 papeles.
   - A cada alumno se le asigna un número entre 1 y 2.500 y se seleccionan generando 100 números al azar de cuatro dígitos utilizando algún programa de computación.
   - Para cada año se asigna a cada alumno un número entre 1 y 500, y se elige 1 de cada 25 alumnos.
   - Se eligen al azar una división de cada uno de los años y se seleccionan 20 alumnos de cada división.
   - Se eligen al azar 60 alumnos de los primeros 3 años y 40 alumnos de los últimos dos años

- Se eligen al azar 60 alumnos de los primeros 3 años y 40 alumnos de los últimos dos años. Se seleccionan en forma separada los varones y las mujeres de acuerdo con la proporción de mujeres y varones que tiene la escuela.

5. En un programa de radio se invitó a las/los oyentes a contestar la siguiente pregunta: "¿Si pudiera volver el tiempo atrás volvería a tener hijos?" De más de 10.000 respuestas el 70% dijo no. ¿Qué muestra esto?

   Elegir, entre las cinco siguientes, la respuesta que mejor responde a esta última pregunta.

   a. La encuesta no dice nada porque arrastra el sesgo por respuesta voluntaria.
   b. No se puede decir nada sin saber las características de los oyentes.
   c. Para sacar una conclusión, es necesario separar las respuestas entre hombres y mujeres.
   d. Hubiese tenido más sentido tomar una muestra aleatoria de las 10.000 respuestas para sacar conclusiones.
   e. Es una muestra legítima elegida al azar entre todos los que escuchan ese programa y tiene un tamaño suficiente como para concluir que la mayoría de los oyentes lo pensarían dos veces antes de tener más hijos.

6. Indicar cuál o cuáles de las siguientes afirmaciones son válidas.

   a. Las respuestas que se obtienen al utilizar un cuestionario expresado en términos no neutrales tendrán "sesgo por respuestas".
   b. Las encuestas de respuesta voluntaria subestiman a la gente con opiniones muy firmes.
   c. Las encuestas de respuesta voluntaria generalmente sobre representan las respuestas negativas.
   d. En general, es posible reducir el sesgo tomando muestras muy grandes, cuanto más grande es el tamaño de la muestra mejor.
   e. El tamaño de la muestra no tiene nada que ver con el sesgo.
   f. Los resultados que se obtienen de un censo son siempre más precisos que los que se obtienen de una muestra, sin que importe cuán cuidadoso haya sido el diseño y su aplicación.

7. Datos - variables

Los datos numéricos son valores de variables numéricas.
Los datos categóricos son valores de variables categóricas.

Las **variables** son **características** que pueden tomar valores diferentes de una unidad a otra, como la edad de las personas, la cantidad de habitantes de cada ciudad, la duración o el consumo de una lamparita.

¿Datos y variables? ¿Son o no son lo mismo?

¿Entonces que son los datos?

Los **datos** son los **valores** observados de las variables.

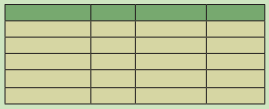
Para ilustrar los conceptos, consideremos la siguiente tabla. Muestra una parte de la libreta donde la maestra registra datos de sus alumnos.

Alumno	Lengua	Matemática	Ciencias Naturales	Participación	Certificado de Vacunas
Cortez María	8,25	6,12	9,51	Buena	Si
García Lobos, Federico	6,59	9,06	8,47	Regular	Si
Gordon, Susana	9,07	7,39	9,72	Buena	Si
Medignone, Horacio	7,55	6,42	8,64	Mala	No
Vázquez, Florencia	6,25	9,63	7,59	Buena	Si

Las unidades son los alumnos del grado, identificados mediante la variable "Alumno" cuyos valores son el nombre y apellido de cada uno de ellos (primera columna de la tabla). Las cinco columnas restantes contienen el **nombre** y los valores de las demás **variables.**

Los nombres encabezan las columnas: Lengua, Matemática, Ciencias Naturales, Participación, Certificado de Vacunas, y en el cuerpo de la tabla (filas a continuación) aparecen los **valores** de cada una de ellas.

Nombres de las variables
Valores observados de las variables (datos)



Las variables tienen un **nombre** y un **valor** para cada individuo de la población.

Los **datos** son los **valores** observados -medidos- **de las variables** para los individuos de una muestra.

Los datos solos dicen muy poco, si no sabemos a qué variables corresponden.

7.1 Variables numéricas y variables categóricas

Los **datos numéricos** son valores de variables numéricas. Los **datos categóricos** son valores de variables categóricas.

En el ejemplo de la libreta de anotaciones de una maestra, las columnas 2, 3 y 4 dan el promedio de notas en cada una de las asignaturas, se trata de **variables numéricas.** La primera, muestra el nombre y apellido de cada alumno; la quinta, el grado de participación en clase registrado en 3 categorías, y la sexta, si la/el alumna/o presentó o no presentó su certificado de vacunas. Todas ellas son **variables categóricas.**

La estadística trata con números, pero **no todas las variables son numéricas.** En este ejemplo, la primera y las dos últimas son **categóricas.** Para resumir los valores de este tipo de variables utilizamos **cantidades y porcentajes.** Por ejemplo, podemos calcular la cantidad de alumnos que se llaman "Juan", o que entregaron el certificado de vacunas, o el porcentaje de alumnas/os que tienen una participación "Buena".

La mayoría de las variables (y por consecuencia también de los datos) se pueden clasificar en **numéricas** y **categóricas.** También se los denominan **cuantitativos** y **cualitativos** respectivamente.

Para analizar variables categóricas se utilizan **cantidades, proporciones** y **porcentajes.**

Ejemplo:

En el censo de población de la República Argentina del año 2001, una de las preguntas fue: ¿Cuál es el grado de educación de las personas con 15 años y más?
La tabla 7.1 responde a esa pregunta. Su título permite ver, inmediatamente, de qué se tratan los datos. Se consigna el año porque estos datos cambian con el tiempo.

Al pie figura, la fuente de los datos: el INDEC. En la primera columna de la tabla se presentan los nombres de las categorías de la variable "Nivel de Educación"; en la segunda y tercera su distribución. En la segunda columna, la distribución se expresa en cantidades, con el encabezamiento indicando "Cantidad de personas". En la tercera columna, la distribución se expresa en porcentajes como también lo muestra su encabezamiento. Suele ser más sencillo pensar en porcentajes. Es más fácil decir el 48,9% tiene estudios primarios completos, que decir que 12.720.081 personas tienen estudios primarios completos.

Distribución del Nivel de educación de la población de 15 años y más. 2001 TABLA 7.1

Nivel de Educación	Cantidad de personas	Porcentaje
Sin instrucción (1)	962.460	3,7
Primario incompleto	3.693.766	14,2
Primario completo	12.720.081	48,9
Secundario completo	6.373.046	24,5
Terciario completo	2.263.082	8,7
Total	26.012.435	100

(1) incluye nunca asistió, jardín e inicial.

Fuente: *INDEC. Dirección Nacional de Estadísticas Sociales y de Población. Dirección de Estadísticas Sectoriales en base a procesamientos especiales del Censo Nacional de Población, Hogares y Viviendas 2001.*

Distribución de una variable: La distribución de una variable nos dice cuáles son sus posibles valores y con qué frecuencia aparecen.

La tabla 7.1 muestra la distribución de la **variable categórica "Nivel de educación"**, máximo nivel de educación alcanzado por las personas de 15 años o más. Tiene 5 categorías: "Sin instrucción", "Primario incompleto", "Primario completo", "Secundario completo" y "Terciario completo". La columna encabezada por "Cantidad de personas" muestra **la frecuencia de cada una de las 5 categorías,** esto es, la **cantidad de personas** que pertenecen a esa categoría. Se trata de **frecuencias absolutas. La suma de las frecuencias** da como resultado la **cantidad total de datos,** 26.012.435, es la cantidad de personas de 15 años ó más en el año 2001.

La frecuencia relativa es el cociente entre la frecuencia absoluta y la cantidad total de datos. Su suma es 1. Cuando las frecuencias relativas están expresadas en porcentaje, la suma es 100, como vemos en la tercera columna de la tabla 7.1.

7.1.1 Gráficos para datos categóricos

7.1.1.1 Gráficos circulares

Utilizaremos un gráfico circular, también llamado gráfico de torta, para visualizar la distribución de la variable "nivel de educación" (tabla 7.1). Podremos visualizar los porcentajes de personas que pertenecen a cada una de las 5 categorías.

Gráfico circular: Se utiliza para representar la distribución de los valores de una variable categórica. El círculo representa el total de los datos. Cada sector dentro del círculo representa una categoría con el ángulo proporcional a su tamaño (cantidad o porcentaje que pertenece a dicha categoría).

Para realizar un gráfico circular, primero se dibuja un círculo. Los 360º representan el total, en este caso todas las personas de 15 años o más de la República Argentina en el 2001. Cada sector dentro del círculo representa una categoría con el ángulo proporcional a su tamaño (cantidad o porcentaje). El sector correspondiente a la categoría "Secundario completo" tendrá un ángulo de 0,245 x 360=88,2 grados.

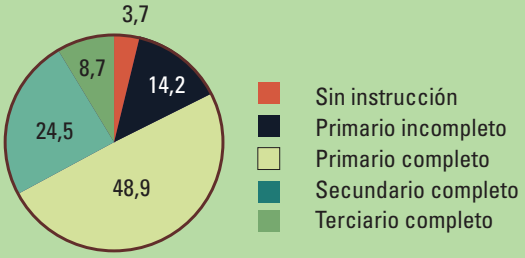


Figura 7.1. *Gráfico circular de la distribución del nivel de educación de las personas de 15 años y más de la República Argentina. Año 2001*

Los gráficos circulares permiten visualizar cómo las partes forman el total, aunque es más difícil comparar ángulos que longitudes. Estos gráficos no son buenos para comparar con precisión los tamaños de las diferentes partes, para eso lo gráficos de barras son mejores.

Los gráficos circulares muestran sectores de área proporcional al porcentaje del total correspondiente a cada grupo o categoría, pero generalmente no muestran la cantidad total en cada grupo, en términos de unidades originales (pesos, número de personas, etc.). Este enfoque se traduce en una pérdida de información.

Para ilustrar esa situación consideremos los datos proporcionados por la Lotería de la Provincia de Buenos Aires en Junio de 2008 http://www.loteria.gba.gov.ar/ sobre como reparte sus ganancias entre diferentes organismos de la provincia.

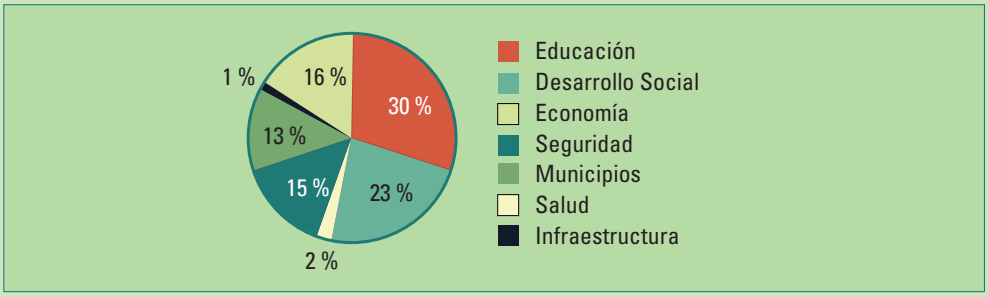


Figura 7.2. *Gráfico circular de la distribución las ganancias de la Lotería de la Provincia de Buenos Aires junio de 2008*

Vemos los porcentajes destinados a los diferentes organismos. Se destinó más del 50% entre Educación y Desarrollo Social. Pero, ¿cuánto fue realmente, en pesos? Veamos esa información en la tabla siguiente.

Siempre se puede pasar de cantidades a porcentajes. En la página de la Lotería de la provincia de Buenos Aires aparecen las cantidades totales y las destinadas a educación por mes, para el período enero-julio de 2008, pero aunque no están los porcentajes podemos calcularlos:

Año 2008	Educación	Total mensual	Porcentaje
Enero	37.307.382	143.225.097	26%
Febrero	45.541.083	164.313.370	28%
Marzo	34.872.907	130.834.379	27%
Abril	32.646.300	116.425.710	28%
Mayo	25.241.707	96.293.288	26%
Junio	35.416.187	117.960.104	30%
Julio	45.553.614	139.475.636	33%

DISTRIBUCIÓN DE LAS GANANCIAS DE LA LOTERÍA DE LA PROVINCIA DE BUENOS AIRES DE JUNIO DE 2008 POR ORGANISMO. TABLA 7.2

Organismo	Junio 2008
Educación	35.416.187
Desarrollo Social	27.370.667
Salud	2.843.829
Seguridad	17.224.945
Municipios	15.141.832
Infraestructura	1.413.519
Economía	18.549.125
Total	**117.960.104**

No se puede ir de los porcentajes a los valores originales sin el conocimiento del total. Esta falta de información puede ser un verdadero problema, por ejemplo, cuando los gráficos muestran los resultados de una encuesta de opinión. Para evaluar el margen de error del porcentaje de personas que respondieron a la pregunta de una manera determinada es necesario saber cuántas personas respondieron la encuesta.

7.1.1.2 Gráficos de barras

Los gráficos de barras se utilizan para representar la distribución de los valores de una variable categórica.

Las categorías se representan en el eje horizontal y la cantidad, o el porcentaje, de datos en el eje vertical. **La altura de las barras** sobre cada categoría representa la cantidad de datos de cada una de ellas. Tal como ocurre con los gráficos circulares, divide a los datos en grupos correspondientes a las categorías y muestra cuántos, o qué porcentaje de individuos pertenecen a cada categoría. Mientras que los gráficos circulares utilizan fundamentalmente porcentajes para indicar el tamaño de cada clase, los gráficos de barras utilizan tanto cantidades como porcentajes.

La figura 7.3 muestra un gráfico de barras de la distribución de los valores de la variable "Nivel de Educación". La altura de cada barra representa los porcentajes de las personas de más de 15 años con nivel de educación mostrado en su base. La barra sobre la categoría "Primario Completo" es la más alta, es la categoría con la mayor cantidad de personas. Podemos comparar categorías: vemos que son más los individuos que tienen el secundario completo, que aquellos que no completaron su educación primaria.

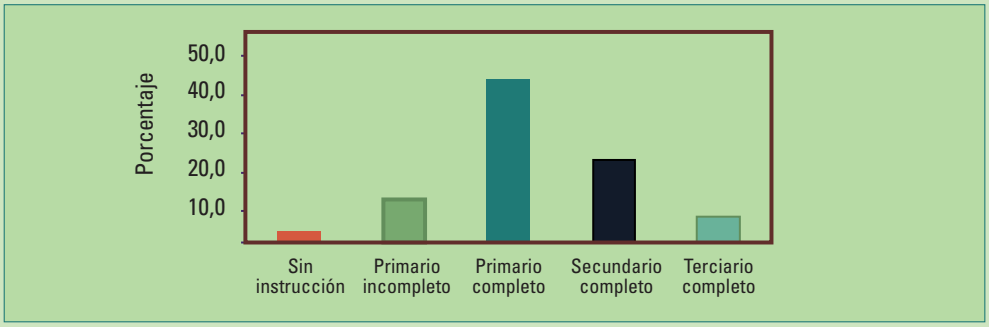


Figura 7.3. *Gráfico de barras de la distribución de la población de 15 años y más de la República Argentina, según máximo nivel educativo. Año 2001*

El gráfico de barras tiene un interés adicional cuando las categorías tienen un orden natural como ocurre en este caso. Vemos que la categoría central "nivel primario completo" es la más poblada y que la caída es más abrupta hacia las categorías correspondientes a menores niveles de educación que hacia los mayores.

Tanto en los gráficos de barras como en los gráficos circulares, los porcentajes de las categorías tienen que sumar 100%:

$$3{,}7\ \% + 14{,}2\% + 48{,}9\% + 24{,}5\% + 8{,}7\% = 100\%$$

7.1.2 Dos variables categóricas

Retomando el tema del nivel de educación, el INDEC incluye los totales y los porcentajes por nivel de educación y género en la presentación de la información de la distribución de la población de 15 años y más de la República Argentina. La tabla nos muestra cómo se distribuyen en forma conjunta dos variables categóricas, nivel de educación y género.

Podemos calcular las cantidades de todas las casillas que nos interesen.

DISTRIBUCIÓN DE LA POBLACIÓN DE 15 AÑOS O MÁS SEGÚN NIVEL DE EDUCACIÓN DE Y GÉNERO. AÑO 2001 TABLA 7.3

Nivel de educación	Total	Total	Género	
			Varón	Mujer
		26.012.435	12.456.479	13.555.956
	Sin instrucción (1)	**3,7%**	3,5%	3,9%
	Primario incompleto	**14,2%**	14,3%	14,1%
	Primario completo	**48,9%**	51,5%	46,5%
	Secundario completo	**24,5%**	23,7%	25,2%
	Terciario completo	**8,7%**	7,0%	10,3%

(1) incluye nunca asistió, jardín e inicial.

Fuente: *INDEC. Dirección Nacional de Estadísticas Sociales y de Población. Dirección de Estadísticas Sectoriales en base a procesamientos especiales del Censo Nacional de Población, Hogares y Viviendas 2001*

A menudo los **gráficos de barras** se utilizan para **comparar dos grupos,** dividiendo la barra de cada categoría en dos y mostrándolas una al lado de la otra.

Un gráfico de barras conjunto nos permite comparar las distribuciones de la variable "Nivel de Educación" en varones y mujeres.

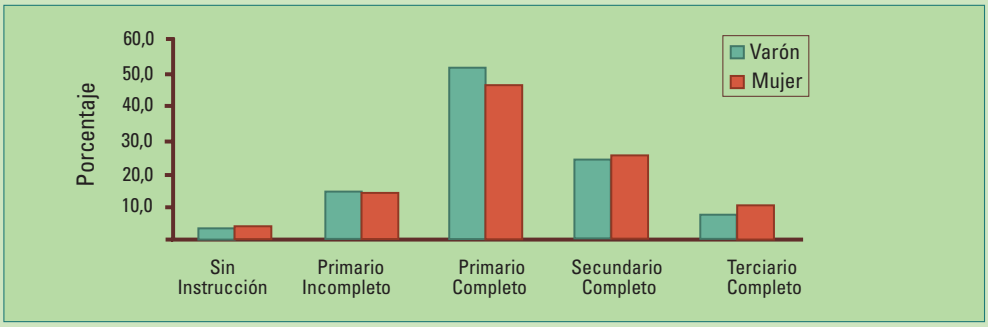


Figura 7.4. *Porcentaje de personas con más de 15 años de acuerdo al nivel de educación y género. Datos tabla 7.3.*

Vemos que en el nivel primario hay más varones que mujeres, pero en el secundario y terciario la relación se invierte, aunque todas las diferencias son pequeñas.

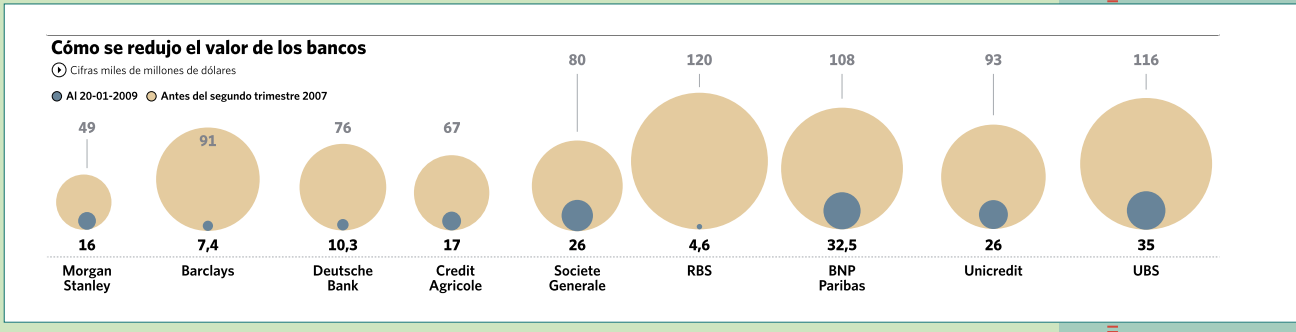
7.2 Actividades y ejercicios

1. Un pictograma es un gráfico de barras que se reemplaza por figuras. Las figuras representan las cantidades o los porcentajes. En forma intencional o no intencional, muchas veces los gráficos exageran las relaciones entre las categorías.

   a. Se utilizó el siguiente pictograma para ilustrar una reducción cercana al 50% de los abandonos de mascotas en la vía pública de una ciudad después de una campaña oficial de concientización. Para reflejar esa reducción sin distorsionar la figura, el artista redujo tanto el alto como el ancho en un 50%:

   Explique por qué la sensación visual de la reducción es bastante mayor que 50%. ¿Cómo debería haber sido la reducción de la figura para reflejarla en forma adecuada?

   b. Un artículo referido a las consecuencias de la crisis financiera de Estados Unidos en 2008 ilustra la reducción de los valores de los bancos mediante el siguiente pictograma. El valor del banco se calcula multiplicando la cantidad total de acciones por su cotización en la Bolsa de Nueva York.



Fuente: *Diario Clarín, 22 de Febrero 2009*

Indique si el pictograma muestra en forma correcta la reducción. Observe que los diámetros de los círculos son proporcionales a los valores.

2. Los siguientes datos son parte de los resultados del primer censo general de la Provincia de Santa Fe (1887). http://www.digitalmicrofilm.com.ar/censos/estadisticas.php

Localización de la vivienda		**Nacionalidades**				**Alfabetización**	
Urbana	90.764	Argentina	92.170	Inglaterra	753	Sí sabe escribir	62.608
Rural	116.712	Italia	46.268	Paraguay	673	No sabe escribir	87.042
Fluvial	2.250	Suiza	5.232	Chile	211		
Otros	382	Francia	2.944	Brasil	192		
		España	2.397	Bélgica	142		
		Alemania	2.070	Portugal	76		
		Austria	1.131	Estados Unidos	74		
		Uruguay	903				

Obtenga un diagrama de barras y un gráfico circular para distribución de los habitantes de la provincia de Santa Fe en 1887 de acuerdo con cada una de las siguientes tres variables categóricas: 1) Alfabetización, 2) Nacionalidades y 3) Localización de la vivienda.

3. Utilice el gráfico que considere adecuado para representar los datos de la tabla siguiente.

Producto Bruto Nominal en dólares per cápita para 10 países de América del Sur, durante 2008 según el FMI

Argentina	8.522	Ecuador	3.927
Bolivia	1.889	Paraguay	2.658
Brasil	8.676	Perú	4.610
Chile	10.814	Uruguay	8.860
Colombia	5.174	Venezuela	11.828

Fuente: *http://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita*

8. Origen de los datos

¿Cuál es el tipo de estudio adecuado para responder una pregunta en particular?

8.1 Censos, encuestas, estudios observacionales y experimentales

Cuando un estudio se basa en consultar toda una población se denomina censo.

Censo: El objetivo de un censo es obtener un registro de todos los miembros de una población en la forma más completa posible. Se relevan las variables principales que permitan la elaboración del marco muestral para futuras encuestas.

Por ejemplo, en un censo de población y vivienda se intenta contactar a todos los habitantes para obtener información respecto de edad, estado civil, género, ocupación y años de escolaridad. Algunas veces, también se recoge información sobre cuestiones relacionadas con las condiciones de vida de la población: en qué tipo de casa viven, si tienen acceso a servicios de salud y a educación, si tienen provisión de agua potable y sistema sanitario, etc. En la República Argentina, desde 1968, el INDEC es el encargado de conducir un censo de población cada 10 años. También realiza el Censo Nacional Agropecuario y el Censo Nacional Económico, con una periodicidad similar.

Un **estudio** puede ser **muestral,** cuando sus resultados están basados en información obtenida a partir una muestra. Este es el caso de la mayoría de los estudios de mercado, encuestas de opinión, evaluación de drogas, etc. En todos estos casos, el objetivo es obtener conclusiones respecto a la población de la que se obtuvo la muestra.

Para muchas investigaciones, provenientes tanto del ámbito privado como público, **no es razonable realizar un censo** por el tiempo y costo que involucra. Más aun, los intentos por recolectar datos completos de una población llevan muchas veces a información de menor calidad.

Un **estudio muestral bien diseñado,** y cuidadosamente conducido, es por lejos superior a un **estudio poblacional** (censo) con **mal diseño** o con escasos recursos. **Si una pregunta está redactada en forma confusa, sus respuestas pueden no tener ningún significado, aunque haya respondido toda la población.**

En una **encuesta** el interés está en **obtener información sobre toda la población estudiando una parte** de ella, es decir una **muestra.** Se intenta recoger información sin perturbar o modificar a la población para no afectar la calidad de los resultados.

Existen numerosos procedimientos para recoger la información a través de un muestreo. Gran parte de la información que nos llega surge de encuestas.

Los estudios pueden ser **observacionales** o **experimentales.**

Todos los **estudios observacionales** comparten un principio: **"sólo mire".** Si le interesa estudiar los hábitos de los pájaros de un bosque y para ganarse su confianza les ofrece migas de pan estará modificando su comportamiento, entonces ya no será un estudio observacional puro.

Las encuestas son estudios observacionales, pero no todos los estudios observacionales son encuestas.

Por el contrario, en un **estudio experimental** se quiere modificar un comportamiento; no sólo observar a los individuos o realizar preguntas sin perturbar. **Se impone en forma deliberada una modificación** de las condiciones para observar las modificaciones generadas.

Por ejemplo, mediante estudios experimentales se podrán responder a las preguntas: ¿Bajaron de peso los alumnos que fueron obligados a duplicar sus horas de actividad física?, ¿Se redujo el índice de mortalidad infantil al habilitar instalaciones de agua potable en una población aislada?

□ 8.2 ¿Pueden estar mal los datos?

Permanentemente, en la televisión, diarios, blogs de Internet, vemos resultados del tipo:

- El 70 % de los que tienen entre 16 y 19 años piensan que bajar música de la red es lo mismo que comprar un CD usado o grabar música prestada de una amiga.
- Para los adolescentes, fumar ya comienza a ser mal visto.
- La depresión es causante de partos prematuros.
- El 49 % de los argentinos tiene sobrepeso.
- En el año 2008 el sueldo de un CEO, el máximo ejecutivo financiero de una corporación (por sus siglas en inglés Chief Executive Officer), era 23 veces más alto que el de un operario, mientras que en el año 1999 lo era 34 veces.

Los buenos datos son el resultado de un enfoque inteligente y un gran esfuerzo. Los datos malos resultan de la falta de cuidado, poco entendimiento del problema o incluso de la intención de producir resultados deliberadamente erróneos. Cuando escuchamos resultados impactantes como los anteriores, lo primero que tenemos que preguntarnos es: ¿con qué calidad de datos se obtuvieron?

□ 8.3 Aspectos éticos

Cuando se recolectan datos de personas, tanto en estudios observacionales como experimentales, surgen complejos aspectos éticos. La situación más complicada se presenta en estudios experimentales en los que se impone un tratamiento a las personas. Los llamados ensayos clínicos son unos de los principales ejemplos. Un ensayo en el que se estudia un medicamento nuevo puede producir, por ejemplo, tanto un daño como un beneficio a los sujetos participantes.

A continuación, describiremos estándares básicos que se deben cumplir en la realización de un estudio que toma datos de personas, ya sea observacional o experimental:

- La organización que lleva adelante el estudio tiene que tener una junta que revise por adelantado los estudios planificados, para proteger a los sujetos participantes de un posible daño. Los hospitales suelen tener un **comité de ética** que se encarga de realizar ese control.
- Antes del inicio, todos los individuos que participan en el estudio tienen que dar un **consentimiento informado.** Tienen que ser informados, con anterioridad a la realización del estudio, sobre la naturaleza del mismo y el riesgo que ocurra algún daño.
    - En una encuesta (estudio observacional) no hay daño físico posible, se debe informar qué **tipo de preguntas** se realizarán y cuánto **tiempo** ocupará responderlas.
    - En los estudios experimentales los sujetos deben recibir la información sobre **la naturaleza y el objetivo del estudio y una descripción de los posibles riesgos.** Luego deben expresar su consentimiento, generalmente por escrito.

El organismo regulador de los ensayos clínicos de la Argentina es el ANMAT (Administración Nacional de Medicamentos y Tecnología Médica, www.anmat.gov.ar)

- Todos los datos individuales se deben guardar en forma confidencial. Solamente se pueden dar a conocer resultados resumidos para grupos de individuos.

□ 8.4 ¿Cómo elegir un tipo de estudio?

¿Cuál es el tipo de estudio adecuado para responder a una pregunta en particular? Por ejemplo, si interesa conocer la opinión de ciertas personas, describir sus estilos de vida y preferencias o describir variables demográficas como nacimientos, muertes o migraciones, es adecuado realizar encuestas, sondeos y otros estudios observacionales. En cambio, si interesa determinar la causa de un resultado o comportamiento (es decir, una razón por la cual sucedió algo), un experimento es mucho mejor. Si no es posible (porque resulta inmoral, demasiado caro, o inviable), la realización de gran cantidad de estudios observacionales - analizando muchos factores diferentes - es la segunda mejor alternativa.

Veremos esto más adelante con mayor profundidad.

8.5 Actividades y ejercicios

1. Indicar y explicar cuál es el tipo de estudio más adecuado para responder a cada una de las siguientes preguntas:

   - ¿Están contentos los alumnos con el nuevo sistema de promoción?
   - ¿El ausentismo de los alumnos es menor en verano que en invierno?
   - ¿El rendimiento de los alumnos en un examen es mejor si durante el mismo escuchan música de Vivaldi, en bajo volumen, en comparación con no escuchar nada?

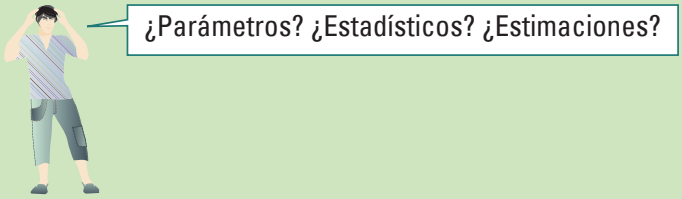
2. Presentar ejemplos de preguntas sobre los estudiantes de una escuela respecto a comportamiento, gustos y opiniones que podrían responderse con cada uno de los siguientes estudios:

   - Una encuesta
   - Un estudio observacional que no sea una encuesta
   - Un experimento

3. Una educadora divide al azar un grupo de niños y niñas de preescolar en dos grupos con iguales capacidades iniciales (para ello les toma una prueba). En un grupo utiliza canciones para enseñarles a contar, y en el otro el método tradicional. ¿Es esto un experimento? Explicar porqué sí o porqué no.

9. “Estadísticos” y “parámetros”

Cuando un estadístico se calcula en base a los datos de toda la población, ese resultado se denomina parámetro.



Aunque la definición parezca nueva, ya nos hemos encontrado con **parámetros** y sus **estimaciones.**

En el ejemplo de las elecciones para presidente del Club Grande de Fútbol (sección 4.2), la verdadera **proporción** de **todos los socios** que están a favor del primer candidato es un **parámetro** que indicamos con la letra p. Describe a la **población de 58.210 socios del club.** Lo llamamos p por proporción, pero no lo conocemos.

La proporción que se obtiene a partir de una muestra es un **estadístico,** lo llamamos $\hat{p}$ (se lee p sombrero).

La investigadora finalmente obtuvo las respuestas de 538 socios, con 274 a favor del primer candidato. La **estimación del parámetro** es:

$$\hat{p} = \frac{275}{538} = 0{,}51$$

El 51% de los socios de la muestra está a favor del primer candidato, lo sabemos porque la investigadora se los preguntó. No sabemos cuál es el porcentaje real de todos los socios que lo apoyan, pero **estimamos** que alrededor de un 51% lo hace.

Consideremos nuevamente la población de todos los socios de un club, pero esta vez observemos su **edad.** El promedio de sus edades es un **parámetro,** lo llamamos **media** de la **variable edad.** Pero si seleccionamos una **muestra** de socios y calculamos el promedio de sus edades obtenemos una **media muestral.** La **media muestral** (capítulo 18) es un estadístico, cuyo valor depende de la **muestra elegida;** se parecerá a la media poblacional (el parámetro) pero en general no será igual.

La media poblacional generalmente se indica por la letra griega mu, **μ.**

Parámetros y estadísticos: Cuando el conjunto de datos proviene de la población completa, el valor del estadístico es un **parámetro.** Un **parámetro** es un número que describe la **población,** pero en la práctica casi nunca sabremos cuál es ese número porque no podemos conocer perfectamente a toda la población.

Cuando el conjunto de datos proviene de una muestra, el número obtenido es el **estadístico** que se utiliza como **una estimación del parámetro.**

La diferencia entre el parámetro y el estadístico es el **error de estimación.**

9.1 Actividades y ejercicios

En cada uno de los siguientes ejercicios

a) Indicar cuál es la unidad muestral, la variable, el estadístico, la población y, cuando corresponda, identificar el tamaño de la muestra.
b) Si el valor en negrita es un parámetro o el valor de un estadístico.

1. Un lote de arandelas tiene un diámetro promedio de **1,908** cm. Este valor se encuentra dentro de las especificaciones de aceptación del lote por parte del comprador. Un inspector selecciona 100 arandelas y obtiene un promedio de **1,915** cm de diámetro. Este valor se encuentra fuera de los especificados límites, por lo tanto el lote es rechazado erróneamente.

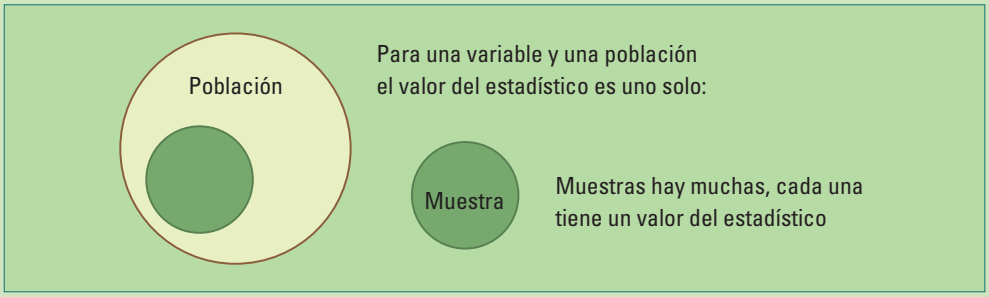
2. En un estudio reciente se entrevistaron 213 familias y la mayoría de las madres estaba al tanto de que los resfríos eran producidos por virus. Pero solamente el **40%** sabía que un antibiótico no puede curar un resfrío, y una de cada 5 creía, en forma equivocada, que un antibiótico lo podía prevenir.

3. En el año 2001 el **50%** de los hogares de la Argentina tenían heladera con freezer, de acuerdo con los valores censales del Anuario Estadístico de la República Argentina de 2006.

4. En el año 2009 el precio promedio de 8 autos modelo 2002 era de **$21.880.**

10. Variabilidad entre muestra y muestra

Siguiendo con el ejemplo del Club Grande de Fútbol, si la encuestadora obtuviera una segunda muestra aleatoria de 538 socios, la nueva muestra estaría compuesta por otros socios (alguno podría coincidir pero muchos serían diferentes). Es casi seguro que no habría exactamente 275 respuestas a favor del candidato 1 como ocurrió con la primera muestra. **Esto significa que el valor del $\hat{p}$ estadístico varía de muestra a muestra:** podría ocurrir que una muestra encuentre un 51% de socios a favor del candidato 1 mientras que una segunda sólo encuentre 37%.



La **primera ventaja** de las muestras aleatorias es que eliminan el **sesgo** del procedimiento de selección de una muestra. Aún así, suele no coincidir el resultado con el verdadero valor, debido a la **variabilidad** que resulta de la selección al azar. Este tipo de variabilidad es llamada **variabilidad muestral.**

Que la **variabilidad muestral** sea muy grande, significa que **el valor del estadístico cambia mucho entre muestra y muestra.** Por lo tanto no podemos creerle al resultado que obtenemos con una muestra en particular. Pero estamos salvados por una **segunda ventaja** que tienen las muestras aleatorias: la variación entre muestra y muestra (de un mismo tamaño) seguirá un patrón predecible. Este patrón predecible muestra que:

Los resultados de muestras de mayor tamaño son menos variables que los resultados de muestras más chicas.

□ 10.1 Muchas muestras

Para ver cuánto le podemos creer al resultado de una muestra debemos preguntarnos ¿qué pasaría si tomásemos muchas muestras de la misma población?

Probemos y veamos en el ejemplo de las elecciones del Club Grande de Fútbol. Supongamos que en realidad (esto no lo sabemos) la mitad de los socios del club (29.105) está a favor y la mitad en contra. Es decir, **la verdadera proporción** (parámetro) de socios que está a favor de uno de los dos candidatos es $p = 0{,}5$

Exactamente el 50% de los socios está a favor del candidato 1.

¿Qué pasaría si utilizáramos una muestra de tamaño 35? Es un tamaño bastante chico para estimar el valor desconocido p de la verdadera proporción de socios a favor del candidato 1.

La figura 10.1 ilustra el resultado de elegir **1.000 muestras diferentes** de tamaño 35 y hallar el valor de $\hat{p}$ para cada una de ellas.

En la primera, de las 1000 muestras, sólo 12 de las 35 personas prefirieron al candidato 1 resultando la proporción $\hat{p} = \frac{12}{35}$

$= 0{,}34$

En la segunda muestra 20 de las 35 personas prefirieron al candidato 1, resultando una estimación de p: $\hat{p} = \frac{20}{35}$ implica que: $\hat{p} = 0{,}57$ para la segunda muestra.

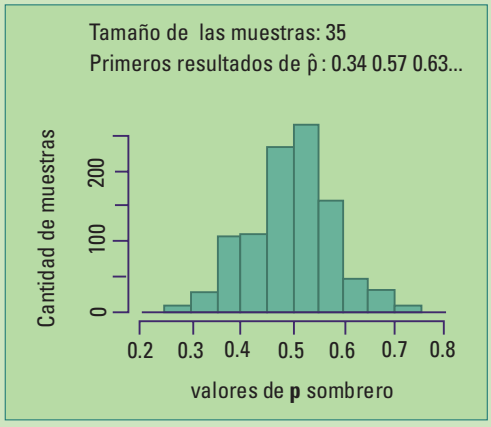


Figura 10.1.

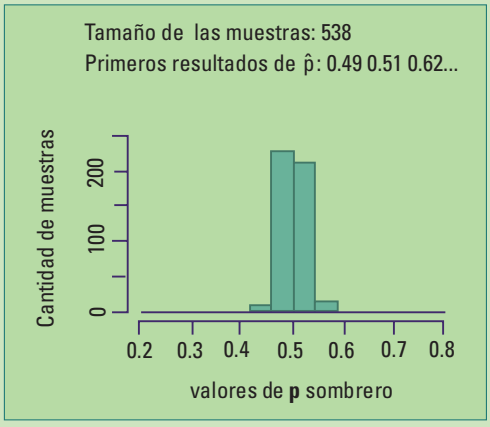


Figura 10.2.

En las figuras anteriores, en el eje horizontal se grafican los valores de las proporciones muestrales $\hat{p}$, las alturas de las barras muestran cuantas de las 1.000 muestras dieron valores dentro de cada uno de los grupos. Este tipo de gráfico se llama histograma (ver capítulo 16).

Vimos, en el capítulo 9, que la investigadora tomó una **única muestra** de 538 socios, y no sólo 35, obteniendo $\hat{p}$=0,51 ¿Qué pasa cuando tomamos muchas muestras de tamaño 538?

La figura 10.2 presenta los resultados de seleccionar **1.000 muestras** aleatorias simples **diferentes,** cada una de **tamaño 538** de una población para la cual la verdadera proporción es p=0,5. Las figuras 10.1 y 10.2 tienen los valores del eje horizontal en la misma escala, esto nos permite ver qué ocurre cuando el tamaño de las muestras se aumenta de 35 a 538: los valores están **más concentrados alrededor del valor verdadero 0,5** y, por lo tanto, **podemos confiar** más en un resultado que proviene de **una muestra de tamaño 538** que de una de tamaño 35.

En ambos casos, los valores de las proporciones muestrales $\hat{p}$ varían de muestra a muestra y están centrados en 0,5. Recordemos que **p=0,5** es el **valor verdadero** del **parámetro.** Algunas muestras tienen un $\hat{p}$ menor que 0,5 y otras mayor, sin que alguno de los dos sentidos esté favorecido. **El estimador de p ($\hat{p}$) no tiene sesgo;** esto ocurre tanto para muestras pequeñas como más grandes.

□ 10.2 Margen de error

Ya vimos que **los valores de $\hat{p}$** que provienen de las muestras de tamaño 35 **están más dispersos** que los de las muestras de tamaño 538 (figuras 10.1 y 10.2). Además, el 95% de las muestras de tamaño 538 dan estimaciones de p entre 0,4592 y 0,5408 o sea 0,0408 a cada lado del valor verdadero 0,5. Llamamos a 0,0408 **margen de error.**

Una **estimación de** p que resulte **de una muestra de tamaño 538** tendrá un **error de a lo sumo 0,0408** en el **95% de las muestras;** sólo el 5% tendrá un error mayor. Decimos que 0,0408 es **el margen de error de la estimación de p** con **un nivel de confianza del 95%.**

Como p=0,5 resulta un margen de error porcentual de 8,16%.

$$\text{Proporción de error: } \frac{0{,}0408}{0{,}5} = 0{,}0816$$

$$\text{Error porcentual: } \frac{100 \times 0{,}0408}{0{,}5} = 8{,}16$$

El margen de error del 8,16% significa que en el 95% de las veces que estimemos el parámetro p con un **tamaño de muestra 538 el error porcentual será menor a 8,16%.**

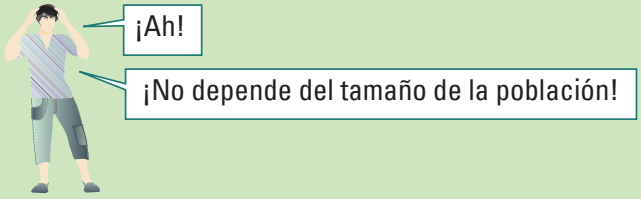
Para las **muestras de tamaño 35 el 95%** de los valores se encuentra entre 0,3429 y 0,6571 dando valores alejados del verdadero hasta una distancia de 0,1571 a cada lado. El **margen de error** porcentual, en este caso sería del **31,42 %.**

Hemos obtenido los márgenes de error tomando muchas muestras de una población a la que le conocíamos el valor verdadero del parámetro p. Este procedimiento es muy incómodo en situaciones reales.

Por suerte, los estadísticos han estudiado el problema de hallar el margen de error en general. Encontraron que cuando se utiliza una proporción muestral $\hat{p}$ calculada a partir de una muestra aleatoria simple de tamaño n para estimar una proporción poblacional p desconocida, con una confianza del 95%, **el margen de error** en un **muestreo aleatorio simple** será aproximadamente de $\frac{1}{\sqrt{n}}$. Al aumentar el tamaño de la muestra se reduce el margen de error.

Cuanto mayor sea el tamaño de la muestra, **MEJOR.**

El margen de error no depende del tamaño de la población, únicamente depende del tamaño de la muestra. Esto es cierto cuando la muestra es sólo una pequeña parte de la población, tal como ocurre en la mayoría de las encuestas. Una muestra aleatoria de tamaño 500 de una población de tamaño 100.000 es tan representativa como una muestra aleatoria de tamaño 500 de una población de tamaño 1.000.000.



Supongamos que se conversa con 20 alumnos/as de una escuela sobre la posibilidad de reducir las vacaciones de verano, agregando dos períodos de vacaciones uno en otoño y otro en primavera. Si a la mitad le pareció una buena idea, ¿estimaría que exactamente la misma proporción de todos los alumnos/as de la escuela está de acuerdo, suponiendo que la muestra es representativa de la opinión de todos?

Si el 50% de la muestra responde sí con una muestra de tamaño:	El porcentaje de la población respondiendo sí podrá ser:	
	Tan bajo como	Tan alto como
10	24%	76%
15	28%	72%
20	**31%**	**69%**
30	34%	66%
50	37%	63%
100	41%	59%
250	44%	56%
1.000	47%	36%

Con una muestra de tamaño 10, si 5 contestaron sí, podría ocurrir que el porcentaje verdadero de alumnos que quieren reducir las vacaciones de verano para agregar dos

vacaciones cortas en otoño y primavera sea tan baja como el 25% o tan alta como el 76%. Esta afirmación es correcta 95 de cada 100 veces. En toda la escuela el porcentaje de alumnos que quieren reducir las vacaciones de verano para agregar dos vacaciones cortas en otoño y primavera se encuentra entre el 24% y el 76%. Es un rango de valores muy amplio para el posible apoyo o no apoyo de la propuesta. Sería conveniente ampliar la muestra para tener un resultado más preciso.

¿Qué significa **"una confianza del 95%"**?

Significa que ese margen de error será válido el 95% de las veces que se calcule el estimador, **confiamos** que nos toque uno de los resultados buenos porque están en una relación de 95 a 5 con los resultados malos.

¿Qué significa **"margen de error"**?

El **margen de error** mide la diferencia máxima que se espera tener entre un resultado obtenido a partir de una muestra y su valor poblacional verdadero, el 95% de las veces.

□ 10.3 Error debido al muestreo aleatorio

Por más que una encuesta esté bien diseñada y bien conducida, dará el valor de un **estadístico** como estimación del **parámetro** poblacional. **Muestras diferentes darán valores diferentes** y el error debido al muestreo estará siempre presente. Pero podremos decir, con cierto grado de confianza, cuál va a ser la magnitud de ese error (denominado margen de error en la sección anterior). Se trata de **errores aleatorios,** surgen de utilizar una muestra en vez de la población total.

□ 10.4 Errores que no son debidos al muestreo aleatorio

Podemos llamar **equivocaciones** a algunos de estos errores. Pueden ocurrir en cualquier encuesta e incluso en los censos. Estas equivocaciones son posibles en todos los pasos, desde el registro del dato hasta obtención final del valor del estadístico. Actualmente, con el uso de procedimientos computarizados para muchos de los cálculos, se han reducido los errores de cálculo.

Otro tipo de errores son los debido a la presencia de sesgos en el muestreo, en las respuestas y/o en su registro (sección 6.3). Por ejemplo, pueden ocurrir cuando un **encuestado miente.** Lo llamamos sesgo de respuesta. Un respondente puede mentir respecto de su edad, de cuántas horas trabaja por día (puede pensar que trabaja poco y entonces las aumenta), de su salario (puede no querer que se sepa que gana mucho, o que gana poco) o puede haber olvidado cuantos paquetes de cigarrillos fumó la semana anterior.

Lo que no mide el margen de error: **No mide** el error que se comete debido al **sesgo** en el muestreo, ni el generado por las respuestas incorrectas y su registro. Estos pueden ser muy grandes, en comparación con el llamado margen de error y **no se reducen** al aumentar el tamaño de la muestra.

Podemos utilizar la analogía del juego del tiro al blanco para describir el efecto del sesgo y el tamaño de muestra en el error de muestreo. Supongamos que el centro del blanco (punto rojo de la figura 10.3) es el parámetro poblacional al que queremos acertar. Si estamos realizando un muestreo aleatorio, en cada muestra -es decir para cada tiro- obtendremos un punto cercano al centro. Algunas veces, el dardo caerá un poco arriba otras un poco abajo. Si en cambio el procedimiento tiene sesgo, los valores estarán todos desviados en una misma dirección. El esquema de la figura 10.3 muestra en la parte inferior puntos negros más concentrados que los de la parte superior, están representando un aumento en el tamaño de las muestras y una reducción de la variabilidad de los resultados. Sin embargo, el error la distancia de los puntos negros al rojo no se reduce al reducirse la variabilidad cuando hay sesgo (parte derecha del esquema).

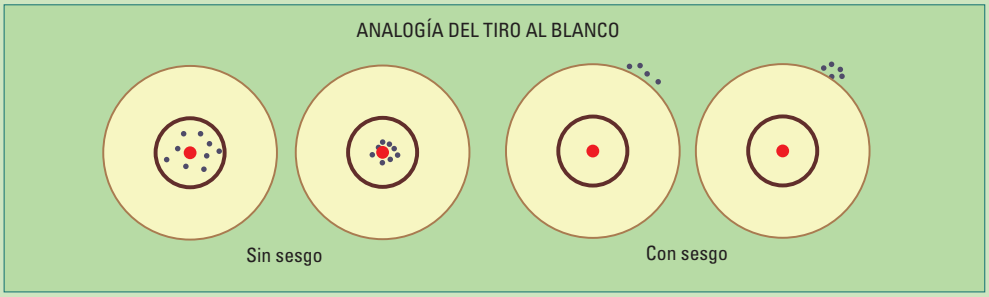


Figura 10.3. *Los puntos del panel inferior están más concentrados, los de la izquierda (representando un muestreo sin sesgo) están más cerca del punto rojo que los de la derecha.*

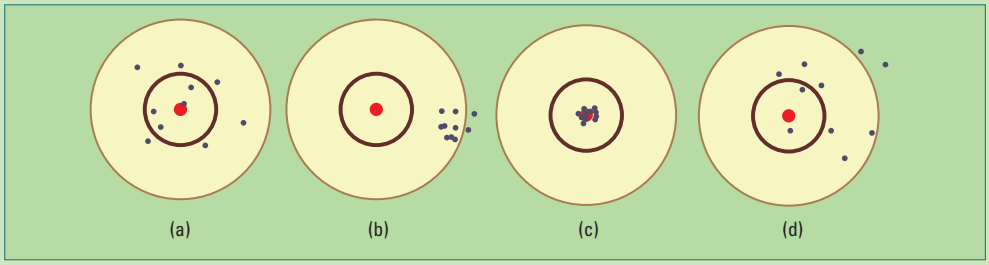
10.5 Actividades y ejercicios

1. Suponiendo que la verdadera proporción de socios a favor del candidato 1 del Club Grande de Fútbol fuera p=0,5.

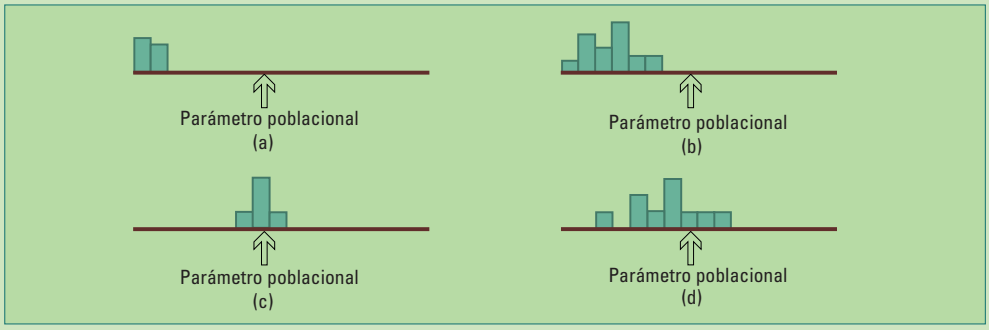
   a) Si el tamaño de la muestra fuera 538,
   b) Si el tamaño de la muestra fuera 35,

   ¿Le sorprendería obtener 51% de socios a favor del candidato 1, y 37%? Para responder utilice los histogramas de 1.000 valores de $\hat{p}$ (figuras 10.1 y 10.2).

2. Siguiendo con la analogía del tiro al blanco de la sección 10.4, indique en cuál de las figuras siguientes los tiros son: I) precisos y sin sesgo, II) precisos y con sesgo, III) imprecisos y sin sesgo, IV) imprecisos y con sesgo.



3. La siguiente figura contiene gráficos semejantes a los de las figuras 10.1 y 10.2 para muchas repeticiones de distintos tipos de muestreos. Las alturas de las barras representan la frecuencia con la que apareció el valor del estadístico. El valor verdadero del parámetro está indicado. Agregue en cada gráfico a qué tipo de muestreo corresponde: I) preciso y sin sesgo, II) preciso y con sesgo, III) impreciso y sin sesgo, IV) impreciso y con sesgo.



4. Se eligen 3 alumnos para representar a su división en el centro de estudiantes. Si su división estuviera compuesta por 15 mujeres y 20 varones, y los representantes seleccionados fueran todos varones ¿Habría que sospechar de discriminación en contra de las mujeres?

   Veamos el comportamiento de la variabilidad muestral al elegir muestras pequeñas de una población pequeña (su división). Escriba los nombres de cada uno de los alumnos/as en papelitos del mismo tamaño y la misma forma. Coloque todos en una bolsa. Luego de mezclarlos, retire de la misma tres papelitos con los nombres de los alumnos seleccionados. Registre la cantidad de mujeres seleccionadas y devuelva los papelitos a la bolsa. Repita 25 veces. Construya un histograma como el de la figura 10.1 para la cantidad de mujeres seleccionadas en las 25 repeticiones. ¿Cuál es la cantidad promedio de mujeres en las 25 muestras?

5. Una encuesta nacional realizada a 437 varones y 1.125 mujeres obtuvo que al 64% de los varones les gustaría ver fútbol femenino por televisión, mientras que ese porcentaje se redujo a 42% entre las mujeres.

   a) Los encuestadores publicaron que el margen de error para una confianza del 95% es aproximadamente del 5% para los varones y 3 % para las mujeres. Explique a qué se debe esta diferencia.

   b) ¿Por qué es necesario incluir el margen de error al dar el resultado de una encuesta?

11. Estudios experimentales

El secreto está en la comparación.
Grupo tratamiento versus grupo control.

Primero, una anécdota.

□ 11.1 La Dama del té

Al preparar un té con leche fría, ¿el sabor es el mismo al verter el té sobre la leche o la leche sobre el té?

Hacia fines de los años veinte (1920) en la ciudad de Cambridge (Inglaterra), una tarde de verano en una reunión de distinguidos académicos y sus esposas, una dama afirmaba: "el sabor no es el mismo". Allí se encontraba Ronald Fisher, quien se entusiasmó en la discusión sobre si era posible saber si la dama podía, realmente, distinguir las dos formas de la preparación del té, únicamente por su sabor. Propuso presentarle varias tazas de té; algunas preparadas con el té vertido sobre la leche y otras con la leche vertida sobre el té. Varios asistentes a la reunión se unieron a la propuesta para ponerla en práctica vertiendo el té y la leche en distintos ordenamientos para que no pudiera adivinar. Así, una a una, fueron ofreciéndole las tazas de té, y registrando la respuesta de la dama sin realizar comentario alguno.

Ronald Aylmer Fisher (1890-1962) Matemático, estadístico, biólogo evolutivo y genetista inglés que estableció los cimientos de la estadística moderna. Statistical Methods for Research Workers (1925), The Genetical Theory of Natural Selection (1930), The design of experiments (1935), Statistical tables (1947).

Al diseñar el experimento se tratan de evitar los aciertos por casualidad. Si se le presenta una única taza y, simplemente adivina, su chance de acertar es 1 en 2. ¿Cuántas tazas son necesarias para reducir los aciertos casuales?

Ronald Fisher incluyó la anécdota en su libro "El diseño de los experimentos en 1935". Mostró experimentos con diferentes diseños, para determinar si la Dama del Té podía detectar la diferencia, así como los cálculos de probabilidades de los aciertos por casualidad.

¿Pero qué pasó con la señora esa tarde? Dicen que acertó el orden de la preparación en todas las tazas.

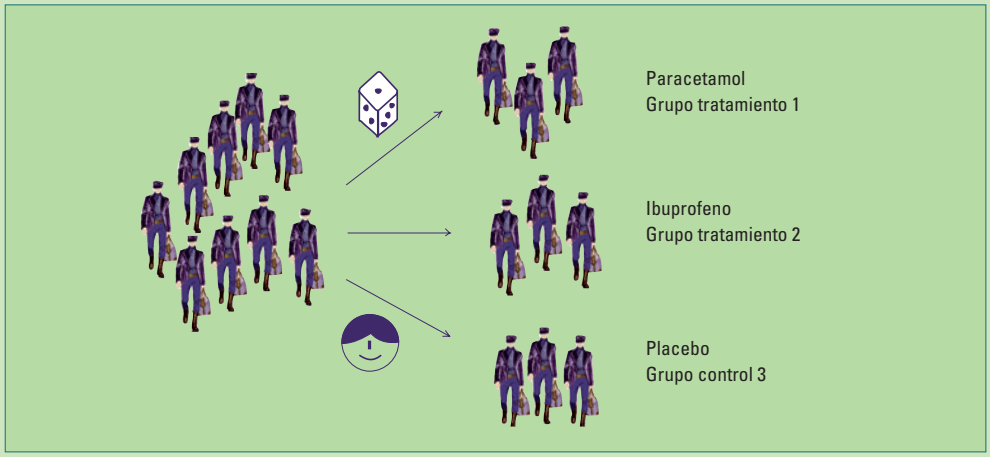
□ 11.2 Vocabulario

Cuando en un estudio se impone un cierto grado de control sobre los participantes y su entorno, como cuando se restringe una dieta o se administra una determinada dosis de un medicamento, se trata de un **estudio controlado.** Los participantes del estudio se asignan a dos o más grupos mediante un mecanismo aleatorio. Cada grupo recibe, por ejemplo, una dosis prefijada de una droga o diferentes drogas alternativas. Construimos así un **experimento comparativo aleatorizado,** llamado también simplemente **estudio controlado.**

El objetivo de la mayoría de los estudios experimentales es señalar una dirección causa-efecto entre dos variables. Se intenta resolver interrogantes del tipo: ¿cuál es la relación entre consumo de alcohol y problemas de visión? También como los siguientes:

- ¿Beber una copa de vino reduce la capacidad para conducir un automóvil?
- ¿Realizar actividad física mejora la fuerza de las mujeres de más de 55 años?
- ¿Tomar suplementos alimenticios con zinc ayuda a reducir la duración de un resfrío?
- ¿La letra con sangre entra?
- ¿La forma y la posición de su almohada modifican la calidad de su sueño?
- ¿La altura del taco de los zapatos influye en la comodidad del pie?

En estudios experimentales es frecuente comparar la efectividad de distintos tratamientos.



Consideremos un estudio comparando dos drogas con un placebo:

- **Participantes del estudio:** pacientes con osteoartritis.
- **Drogas:** acetaminofeno (también llamado Paracetamol o Tylenol), ibuprofeno.
- **Variable Respuesta:** grado de reducción del dolor de la rodilla y la cadera.

Los participantes se asignan a uno de tres grupos aleatoriamente. Dos grupos se llaman de tratamiento y uno grupo llamado de control. Los dos **grupos tratamiento** reciben Paracetamol e Ibuprofeno respectivamente y el **grupo control** recibe un placebo. En este estudio se pueden realizar preguntas como: ¿en cuál de los grupos la reducción del dolor ha sido la mayor? o ¿en cuál de los grupos se registra una menor proporción de eventos indeseables como, por ejemplo, gastrointestinales?

El experimento es aleatorizado cuando los sujetos han sido asignados a los distintos grupos mediante un mecanismo aleatorio.

El experimento es comparativo cuando sus conclusiones se obtienen comparando los resultados de los distintos grupos. Se espera que los grupos sólo difieran en la característica estudiada.

11.2.1 Grupo tratamiento versus grupo control

El grupo, o los grupos, tratamiento se componen de pacientes que reciben algún tratamiento. En nuestro ejemplo, teníamos dos grupos tratamiento uno asignado a Paracetamol y el otro a Ibuprofeno. En general, el grupo control se compone de individuos no tratados, o que reciben un tratamiento estándar bien conocido, cuyos resultados se compararán con el tratamiento nuevo.

11.2.2 Placebo

Un placebo es un tratamiento falso e inocuo, como una píldora de azúcar. A menudo se da a los miembros del grupo de control, para ocultar si están tomando el tratamiento (por ejemplo, Paracetamol) o no están recibiendo ningún tratamiento en absoluto.

El placebo es dado a los participantes asignados al grupo de control, precisamente con el fin de controlar el fenómeno llamado **efecto placebo:** los pacientes informan algún tipo de efecto cuando tienen la percepción de estar realizando un tratamiento, como tomar una píldora (aunque sea una píldora de azúcar).

El efecto informado puede ser positivo “Sí, me siento mejor”, o negativo “Me estoy sintiendo mareado”. Sin un placebo como referencia para hacer comparaciones, los investigadores no podrían distinguir si los resultados son debido al efecto real del tratamiento o al efecto placebo.

11.2.3 Ciego, doble ciego y triple ciego

En un **experimento ciego los participantes del estudio no saben si están en un grupo de tratamiento o en un grupo de control.** Un experimento a ciegas intenta eliminar cualquier sesgo de respuesta del sujeto debido a la información que recibe. Un paciente puede sentir una mejoría simplemente por saber que está tomando un medicamento de última generación muy bueno, o sentirse mal por saber que el medicamento es nuevo y no ha sido probado aún.

Un **experimento doble ciego** controla los posibles prejuicios tanto de los pacientes como de las/os investigadoras/es. **Ni los pacientes, ni las/os investigadoras/es conocen qué sujetos recibieron el tratamiento y cuáles no.** Un doble-ciego es mejor; los investigadores pueden tener un especial interés en los resultados (por algo están haciendo el estudio).

En un **experimento triple** ciego ni los pacientes, ni las/os investigadoras/es, ni las/os profesionales estadísticas/os que realizan el análisis de los datos pueden identificar a los sujetos con y sin tratamiento. Esto es lo mejor.

12. Estudios observacionales

Es necesario observar el mundo para intentar entender su funcionamiento.

La observación es el primer paso, a partir de ella se abrirán diversos caminos para desarrollar nuevas teorías y modelos; podrá motivar la realización posterior de nuevos estudios. Desde un punto de vista estadístico, los mejores resultados provendrán de estudios experimentales en comparación con estudios observacionales, pero algunas veces sólo es posible realizar estos últimos.

□ 12.1 Observar es bueno

Observando la naturaleza, Charles Darwin descubrió la selección natural como mecanismo de evolución de las especies.

En 1831 a los 22 años, emprendió un viaje alrededor del mundo de 5 años de duración como naturalista sin sueldo en un barco británico de reconocimiento, el velero Beagle. El primer indicio real respecto de la evolución de las especies no fueron los pinzones de las Islas Galápagos (1935), como se afirma muchas veces. Fue **tres años antes en la costa Argentina.** Ancló cerca de Bahía Blanca durante el primer año del viaje; desde allí llegó a Punta Alta y Monte Hermoso donde desenterró restos de fósiles de diversas criaturas, entre ellas encontró especies extintas ligeramente diferentes a las vivas. Le llamó la atención la presencia de un ñandú grande en las pampas y uno más pequeño el sur del río Negro.

A partir de las observaciones realizadas durante ese periplo, ya hacia 1838 Darwin tenía claro cómo la selección natural era un mecanismo de la evolución, aunque demoró la publicación de sus obras.

Consciente de las posibles repercusiones, y del rechazo de esa nueva visión de la realidad biológica por la conservadora sociedad victoriana, postergó su publicación, y decidió continuar añadiendo ocasionalmente nuevos datos.

"El Origen de las especies por selección natural" se puso a la venta recién a fines de 1859, agotándose ese mismo día. En enero de 1860 salió la segunda edición, llegó a tener seis ediciones en total durante la vida de Darwin.

La evolución es, 150 años después de su descubrimiento, tan firme como la "teoría" heliocéntrica (la Tierra gira alrededor del Sol) que también se desarrolló observando sin prejuicios (Copérnico, 1543). Cada una de estas teorías da una explicación confirmada, hasta cierto punto, por medio de la observación y la experimentación. A eso se refieren los científicos cuando hablan de una teoría.

Como ocurre con frecuencia con los avances de la ciencia, Darwin no fue el único en darse cuenta. También lo hizo Alfred Wallace, en forma independiente y simultánea; sus trabajos fueron presentados conjuntamente el 1 de julio de 1858, en la Linnean Society de Londres.

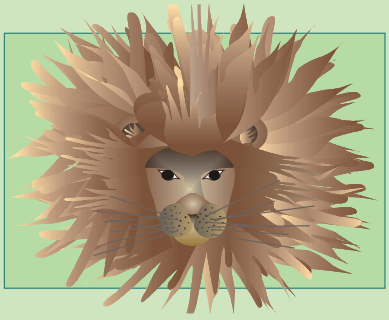
Las prácticas observacionales abren el camino para realizar experimentos. Con la evolución no es fácil experimentar porque en general se manifiesta luego de muchas generaciones.

Una manera de superar esta dificultad consiste en realizar experimentos utilizando especies con ciclos cortos de vida, por ejemplo, bacterias (500 generaciones en 75 días). Se las cultiva alterando alguna condición ambiental para observar la respuesta evolutiva. También se realizan experimentos con organismos superiores como moscas del género Drosophila; completan su ciclo en sólo 12 días, permitiendo detectar cambios generacionales en lapsos cortos y así estudiar su evolución.

Todos estos estudios requieren de la aplicación de técnicas estadísticas para obtener conclusiones con valor científico.

□ 12.2 Cuando sólo se puede observar

Imaginemos una investigación para conocer el comportamiento de los leones, en particular cómo las leonas enseñan a sus cachorros a cazar. Comienza por la observación. Al principio puede ser difícil saber qué registrar. Eventualmente, pueden surgir algunos patrones orientando las mediciones.



- ¿Con qué frecuencia cazan los leones?
- ¿Lo hacen los machos solos?
- ¿Lo hacen las hembras?
- ¿Van en grupos?
- ¿Los acompañan las crías, a partir de qué edad?
- ¿En qué etapa de la caza incorporan a los cachorros?

Las observaciones bien diseñadas, dirigidas a variables definidas claramente permitirán obtener resultados más convincentes.

En muchas oportunidades, no es ético realizar un estudio experimental. Por ejemplo, no es posible forzar a 100 personas a fumar 3 paquetes de cigarrillos por día y a otras 100 uno. En humanos sólo pueden realizarse estudios observacionales para responder preguntas como: ¿fumar provoca cáncer de pulmón?
Para evitar estas dificultades se conducen experimentos con animales, pero cada vez hay más reacciones contra este enfoque.

En un **estudio observacional** se registran algunas características de individuos tratando de no influir en dichas mediciones.

Por ejemplo, se pueden considerar dos grupos de individuos, uno de sedentarios y otros de deportistas, -y sin influir en sus hábitos- se mide su nivel de colesterol en sangre, para evaluar si la actividad física lo afecta.

13. Estudio observacional versus estudio experimental

Algunas veces es posible realizar cualquiera de los dos tipos de estudio. En ese caso ¿cuál elegiríamos?

A continuación describiremos un ejemplo con ambas alternativas. Interesa estudiar si un suplemento diario de calcio en la dieta beneficia a las mujeres aumentando su masa ósea.

Diseño 1:
Se forma un primer grupo seleccionando consumidoras habituales de suplementos de calcio, y un segundo grupo con mujeres, también consumidoras de suplementos, pero sin calcio. Se mide la masa ósea en ambos grupos y se comparan los resultados. Se trata de un **diseño observacional** porque las mujeres **eligen libremente** tomar o no tomar suplementos de calcio.

Diseño 2:
Se selecciona un grupo de mujeres para participar del estudio. A la mitad de las mujeres, **se les asigna en forma aleatoria,** suplementos de calcio, a la otra mitad placebos con el mismo aspecto. Ni el médico ni la participante saben si ella pertenece al grupo de tratamiento o al grupo de control. Después de un tiempo de seguimiento del estudio, se comparan los dos grupos respecto a su masa ósea. Se trata de un **diseño experimental** porque las participantes son asignadas al azar en los grupos.

El enfoque experimental es más adecuado porque, por ejemplo, las mujeres que toman suplementos de calcio voluntariamente podrían ser precisamente las que mejor se cuidan en general y, por lo tanto tener mayor masa ósea por otras razones **(variables de confusión).** Con el diseño experimental, administrando en forma aleatoria el calcio a la mitad de las mujeres, se espera obtener grupos balaceados respecto de las variables que pueden afectar los resultados.

Como vimos en el capítulo 12, algunas veces no es posible realizar estudios experimentales. En estudios observacionales se puede controlar el efecto de los factores de confusión realizando las comparaciones en subgrupos más pequeños y más homogéneos. Por ejemplo, en un estudio sobre el efecto del tabaquismo en la salud, se comparan fumadores con no fumadores dentro de subgrupos con edad similar, el mismo género y características similares en todos los posibles factores influyentes en la patología en estudio además del tabaquismo. Se trata de lograr homogeneidad dentro de los subgrupos excepto por la condición en estudio, en este caso el tabaquismo.

13.1 Actividades y ejercicios

1. ¿Cuáles de las siguientes afirmaciones son verdaderas?

   a) En un estudio experimental un grupo es forzado a seguir un tratamiento con el propósito de observar una respuesta.
   b) En un estudio observacional se recoge información sin realizar ninguna acción para modificar la situación existente.
   c) Las encuestas son estudios observacionales, no son experimentos.

2. ¿Cuáles de las siguientes afirmaciones son verdaderas?

   a) En un experimento los investigadores deciden como se colocan a las personas en los distintos grupos.
   b) En los estudios observacionales, los participantes eligen en qué grupo estar.
   c) Un grupo control generalmente es de elección voluntaria.

3. En un estudio para determinar el efecto de la actividad física en el nivel de colesterol, se comparó el nivel de colesterol de 100 sujetos que concurrían al gimnasio 4 veces por semana con 100 sujetos que no realizaban actividad física. En un segundo estudio, 50 sujetos fueron asignados aleatoriamente para asistir al gimnasio 4 veces por semana y otros 50 para participar en clases de pintura. Indique cuales de las siguientes afirmaciones son verdaderas.

   a) El primer estudio es un experimento controlado, el segundo es un estudio observacional.
   b) El primer estudio es un estudio observacional, el segundo es un experimento controlado.
   c) Ambos son experimentos controlados.
   d) Ambos son estudios observacionales.
   e) Cada estudio es un poco experimental y un poco observacional.

4. En un estudio para determinar el efecto de la provisión gratuita de leche a los niños de las escuelas de un distrito, se asignó a las escuelas en forma aleatoria uno o dos litros de leche por semana por alumno, y se registraron los días de ausencias por enfermedad durante un año. En otro estudio realizado en un hospital de niños se preguntó, mediante un cuestionario entregado en la sala de espera, cuánta leche tomaba el niño por semana y cuántos días había faltado a la escuela por enfermedad en el último año. Indique cuáles de las siguientes afirmaciones son verdaderas.

a) El primer estudio es un estudio experimental sin grupo control y el segundo es un estudio observacional.
b) El primer estudio es un estudio observacional y el segundo es un estudio experimental controlado.
c) Ambos estudios son observacionales.
d) Ambos estudios son experimentos controlados.

5. Se seleccionaron al azar 10 de 20 personas que sufrían dolor de cabeza y se les dio chocolates con sabor a menta y color modificado para ocultar el chocolate. A los otros 10 se les dio tabletas de aspecto y gusto similar pero sin chocolate. Al día siguiente 6 de las 10 personas que habían consumido chocolate reportaron haber sufrido dolor de cabeza; ninguna de las que no recibieron chocolate reportaron dolor de cabeza.

a) Se trata de un estudio observacional para determinar el efecto del chocolate sobre el dolor de cabeza.
b) Se trata de una encuesta en la cual se eligieron 10 de 20 personas con dolor de cabeza para darles tabletas de chocolate con sabor a menta.
c) Se trata de un censo de 20 personas que suelen tener dolor de cabeza; se registró a cuántas personas se les dio chocolate y cuántas tuvieron dolor de cabeza.
d) Se realizó un estudio utilizando el chocolate como placebo para estudiar una causa del dolor de cabeza.
e) Se realizó un experimento en el cual al grupo tratamiento se le dio chocolate y al grupo control no.

6. Indique cuales de las siguientes afirmaciones son verdaderas.Interesa saber cuántos varones y cuántas mujeres asisten a una escuela determinada. ¿Cuál es la forma más adecuada de obtener esa información? Mediante un

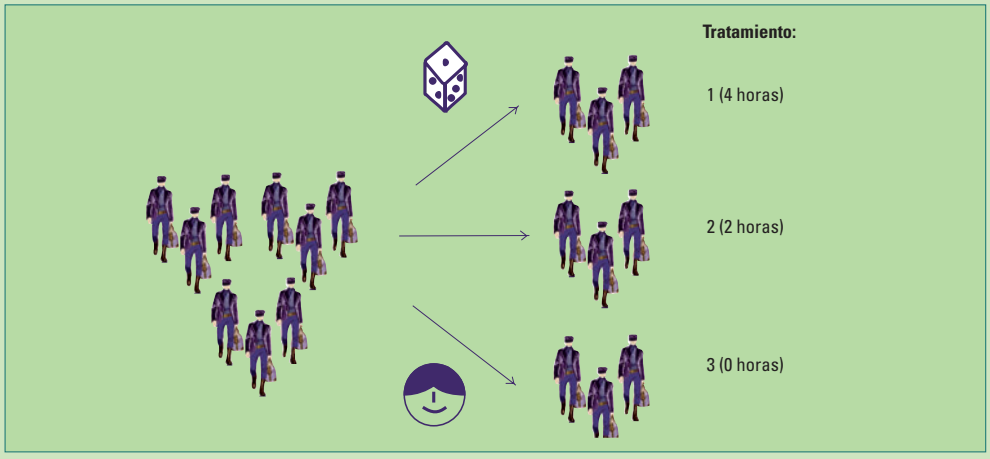
a) Censo
b) Encuesta
c) Experimento controlado
d) Estudio observacional

14. No siempre los tratamientos son tratamientos

En jerga estadística los tratamientos no se usan no sólo en medicina.

Cuando en un estudio interesa el efecto de un único factor **(variable explicativa)** sobre una **variable respuesta** se denomina **tratamiento** a cada uno de los **niveles del factor.**

Comencemos con un ejemplo: interesa evaluar si la asistencia a clases de apoyo los días sábados **(el factor)** puede influir en el rendimiento de los estudiantes **(la respuesta).** Para ello se puede **dividir a los alumnos** seleccionados en **tres grupos.** El primero con 4 horas por semana de clases de apoyo (tratamiento 1), el segundo dos horas (tratamiento 2) y el tercero ninguna (tratamiento 3). Al final del trimestre los alumnos darán una prueba con el fin de comparar las respuestas a los tratamientos.



Es importante que la asignación de los alumnos a los grupos se haga al azar y no en forma voluntaria. El experimento será ciego si los evaluadores ignoran a qué grupo pertenece cada alumno, y nunca será doble ciego por que los alumnos siempre sabrán cuantas horas toman de clases.

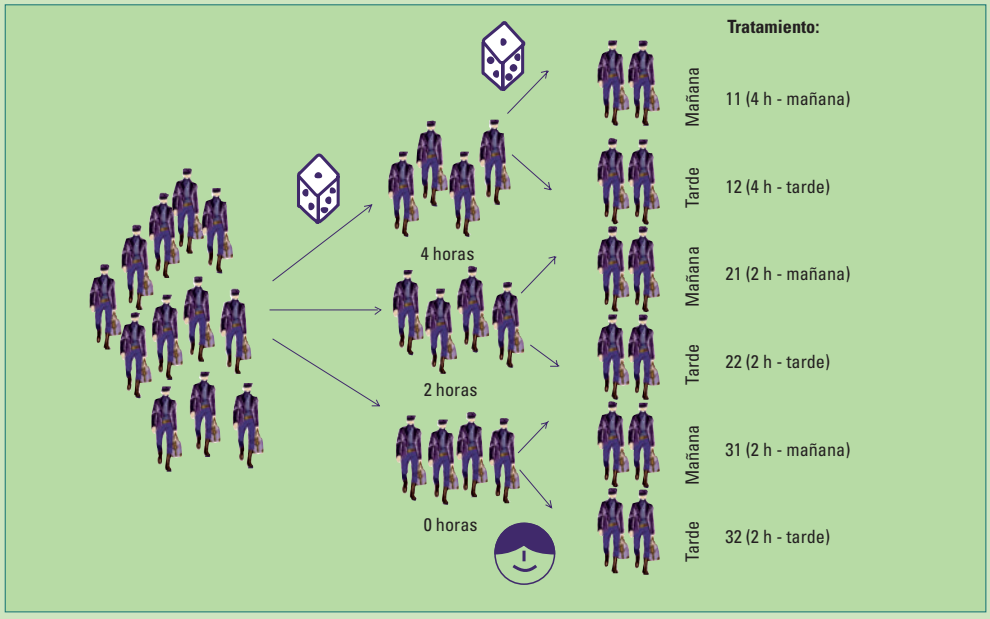
En el ejemplo, el factor es la asistencia a clases de apoyo medido en horas, y la variable respuesta el desempeño del alumno al final del trimestre. Tenemos tres tratamientos, uno para cada nivel del factor (4 horas, 2 horas, 0 horas).

Estamos utilizando el término tratamiento en sentido amplio. En estadística nos referimos a tratamientos no solo a los de medicina, sino también en referencia a distintas áreas, cuando se quiere comparar el efecto de cierto tipo de intervención sobre alguna respuesta. Más precisamente se trata de factores (variables explicativas) que podrían tener efecto sobre variables de respuesta. Un grupo se "trata" con algún nivel de la variable explicativa y el resultado a medir es de la variable respuesta.

Compliquemos un poco el ejemplo. Supongamos que los alumnos fueran a su vez divididos, aleatoriamente, en clases de apoyo en dos turnos: mañana y tarde.

Cuando hay más de un factor en estudio, se denomina tratamiento a cada una de las combinaciones de los diferentes niveles de cada uno de los factores. En este caso, como tenemos dos factores, uno con tres niveles (4, 2, y 0 horas) y otro con dos (mañana y tarde), resulta un total de 6 tratamientos:

- Tratamiento 11: 4 horas - mañana
- Tratamiento 12: 4 horas - tarde
- Tratamiento 21: 2 horas - mañana
- Tratamiento 22: 2 horas - tarde
- Tratamiento 31: 0 horas - mañana
- Tratamiento 32: 0 horas - tarde



En el ejemplo sobre la efectividad de las clases de apoyo se ilustra el uso del término "tratamiento". Es un estudio experimental; pero, este concepto puede aplicarse también a estudios observacionales. Cuando se comparan sujetos que concurren habitualmente al gimnasio 4 veces por semana con sujetos que no realizan actividad física, tenemos un estudio observacional; el factor en estudio es la actividad física y los "tratamientos" son 2:

- concurre 4 veces por semana al gimnasio
- no realizan actividad física.

15. Mediciones válidas

Una variable es una medida válida de un concepto si lo representa adecuadamente.

Medir una característica de un individuo (persona, objeto, animal, etc.) significa asignarle un número expresable en distintas unidades que la represente. El resultado de esa medición es **variable** y toma diferentes valores dependiendo del individuo a quien se le está realizando la medición.

Algunas características, como el peso y la talla, pueden ser más sencillas de medir que otras como la inteligencia o la percepción del dolor.

Muchas veces disponemos de un **instrumento** para realizar la medición. Para obtener la longitud de una mesa utilizamos una cinta métrica; estará expresada en centímetros para la UE y Argentina y en pulgadas para Estados Unidos, es decir, como unidades para expresar mediciones se pueden utilizar centímetros ó pulgadas.

La medición requiere de:

- Un proceso previo de **transformación de conceptos** (longitud, desempleo, dolor, nivel socioeconómico, etc.) en variables definidas con precisión.
- La elección del **instrumento** para medirlas.

La utilización de una cinta métrica para transformar la idea "longitud" en un número es directa, porque sabemos exactamente qué queremos decir con longitud, pero en otros casos puede resultar mucho más complicado. Para medir la inteligencia se requiere de un cuestionario y un mecanismo de cálculo para obtener un número de acuerdo con las respuestas.

Muchas veces no disponemos de mecanismos o instrumentos de medición adecuados para abordar un tema de nuestro interés, pero podemos utilizar resultados de organismos del estado o empresas privadas. Al utilizarlos es importante evaluar cómo están definidos, y si se trata de medidas válidas para describir las propiedades que se pretenden medir.

En la próxima sección presentamos la cantidad de accidentes de tránsito y la cantidad de desocupados como ejemplo de dos conceptos definidos precisamente para obtener, sin demasiadas dificultades, mediciones válidas. Veremos más adelante que a veces no es posible obtener mediciones claramente válidas (dolor, inteligencia). También, que para un mismo concepto se puede obtener más de una medición válida. Finalmente, ilustramos con más detalle la construcción de un instrumento para medir la evolución de los precios al consumidor.

15.1. Sin demasiadas dificultades

15.1.1. Accidentes de tránsito

¿Cómo se mide la seguridad en las rutas? Se puede contar la cantidad de víctimas fatales por año en el momento de un accidente de tránsito. En nuestro país el Registro Nacional de Antecedentes de Tránsito (ReNAT) publica esa información. Como vimos en la capítulo 3, se puede utilizar esa **cantidad de muertes** como **variable** para medir la seguridad en las rutas; también la tasa de víctimas fatales por cada millón de habitantes o por cada cien mil vehículos circulantes.

Podríamos utilizar los datos ReNAT sin averiguar de qué manera se elaboran. Sin embargo, como consumidores de la información, deberíamos indagar un poco más. Para contar muertes fatales en las rutas es necesario saber exactamente a qué se refiere el término "víctimas fatales". ¿Se trata de peatones atropellados por un auto?, ¿automovilistas arrollados por un tren? Contestar estas y otras preguntas, permite saber qué se está contando. ¿Se incluyen los fallecidos dentro de las 24 h del accidente, o dentro del primer mes, etc.?

15.1.2. Desocupación

La Encuesta Permanente de Hogares (EPH, INDEC) releva información para calcular trimestralmente la tasa de desocupación.

Para ser desocupado, es necesario formar parte del mercado laboral e integrar la población económicamente activa .La población económicamente activa es la que se cuenta en el denominador. Entre ella los que están buscando trabajo van en el numerador, queda claro en la definición. Esta población está formada por las personas con alguna actividad económica o que sin tenerla la están buscando activamente. Los estudiantes o los jubilados, no deben contarse como parte de los desocupados aunque no tengan empleo, porque no están disponibles para realizar un trabajo.

La EPH define la población desocupada como el conjunto de todas las personas que, no teniendo ocupación, están buscando activamente trabajo. Este concepto no incluye a los desocupados que han suspendido la búsqueda por falta de oportunidades visibles de empleo, ni a los subocupados involuntarios. Es decir, una persona para ser desocupada debe estar disponible para un trabajo y buscando uno. La tasa de desocupación daría diferente si se utilizara otra definición.

Tasa de desocupación: Es la relación entre la población desocupada y la población económicamente activa, expresada en porcentaje.

$$\text{tasa de desocupación} = 100 \times \frac{\text{cantidad de personas desocupadas}}{\text{cantidad de personas económicamente activas}}$$

Importan mucho los detalles. Los resultados también dependen de cómo se realizan las preguntas para obtener la información relacionada con la desocupación. No se trata simplemente de preguntarle al entrevistado: " ¿Forma parte del mercado laboral?" "¿Está desocupado?"

Se necesitan muchas preguntas para clasificar a una persona en empleada, subempleada o no perteneciente al mercado laboral. De eso se encarga la EPH. Ahora veamos algunos de sus resultados. En particular consideraremos **la tasa de desocupación desde 1995 hasta 2008** y la presentaremos en un **gráfico de tiempo.**

> **Gráfico de tiempo:** Este tipo de gráfico se utiliza para examinar la evolución a lo largo del tiempo de alguna variable. Tiene una unidad de tiempo en el eje horizontal (como meses o años) y en el eje vertical alguna cantidad (ingresos de los hogares, tasa de natalidad, ventas totales, porcentaje de la gente en favor del presidente, y así sucesivamente). En cada período de tiempo, la cantidad está representada por un punto, y los puntos están conectados por líneas.

Los datos correspondientes a la tasa de desocupación deben dividirse en dos períodos. El primero de 1995 hasta 2002 - porque el índice se publicaba 2 veces al año (en mayo y octubre)- y el segundo de 2003 hasta el 2008, porque la publicación se realiza 4 veces al año. En este segundo período, con inicio en enero de 2003, la EPH introdujo mejoras metodológicas. El INDEC **cambió el instrumento de medición** del mercado laboral para mejorar la calidad de la información.

La figura 15.1 muestra la evolución de la tasa de desempleo con un quiebre entre la medición de octubre de 2002 y la del 1er trimestre del 2003, poniendo de manifiesto el cambio de la metodología, Las mediciones de los dos períodos no son comparables directamente.

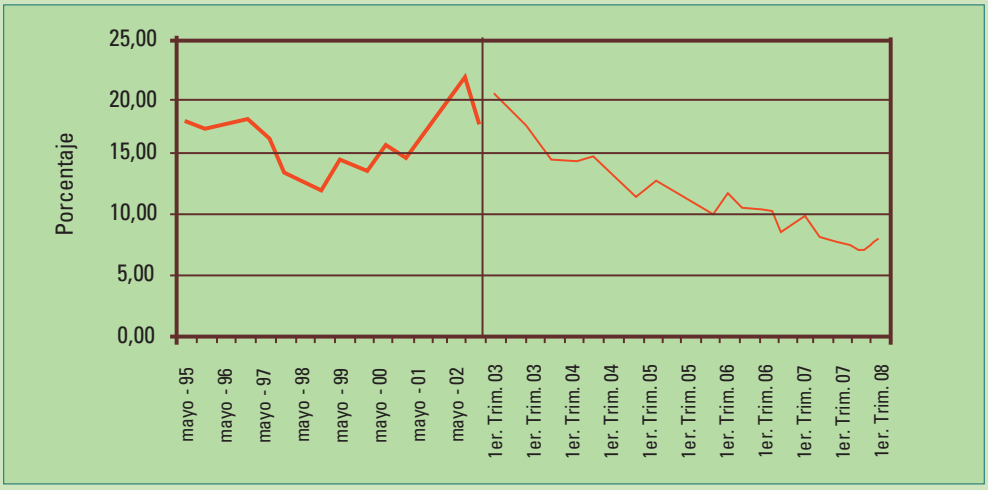


Figura 15.1. *Tasa de desempleo según la EPH desde mayo de 1995 hasta octubre de 2002 (con periodicidad bianual, mayo, octubre) y desde el 1er trimestre del 2003 hasta el 1er. trimestre de 2008 (con periodicidad trimestral).*

Vemos una tendencia decreciente de la desocupación desde mayo de 1995 hasta octubre de 1998, a partir de allí la tasa de desocupación empieza a aumentar llegando a un máximo en mayo de 2002. Luego se observa una tendencia decreciente en la tasa de desocupación hasta el 1er. trimestre de 2008.

Es importante destacar que la escala temporal en los dos períodos no es la misma. No son visualmente comparables las tendencias crecientes y decrecientes entre ellos.

Los cambios metodológicos incluyen mejoras de los formularios, ampliación de la muestra, aumento de la frecuencia con que se recoge la información, incorporación de procedimientos digitales, etc.; son habituales en todos los institutos de estadística del mundo. La comparación de los resultados obtenidos con metodologías diferentes es más difícil.

La figura 15.2 muestra un gráfico de tiempo para la tasa de desocupación de USA desde 1979 hasta 1996. Aquí vemos tres cortes, porque en 1986, 1990 y1994 se produjeron cambios metodológicos en el relevamiento y procesamiento de la información (http://www.bls.gov/cps/eetech_methods.pdf).

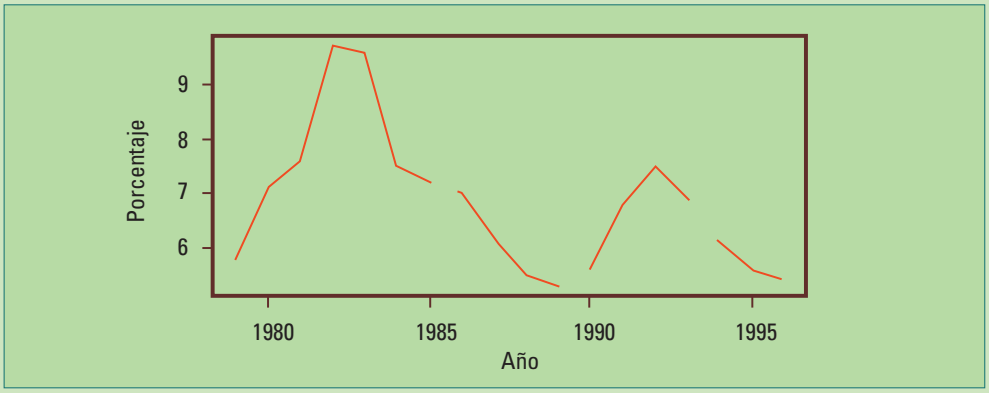


Figura 15.2. *Tasa de desempleo de USA ftp://ftp.bls.gov/pub/suppl/empsit.cpseea1.txt.*

□ 15.2. Puede ser más difícil

Nadie pondría objeciones en el uso de una cinta métrica para medir la longitud de una mesa, sin embargo, algunas personas se pueden oponer a exigir una prueba de evaluación para decidir si un alumno está capacitado para ingresar a una universidad. Aunque todos coincidirán en que es una mala idea medirles la altura a los aspirantes y aceptar a los más altos. ¿Por qué? Porque la altura no tiene nada que ver con estar o no estar preparado para la Universidad.

Una **variable** es una medida **válida** de un concepto si lo representa adecuadamente, o es una característica importante.

¿Está midiendo lo que le interesa medir? Si es así la variable es válida; si no, no lo es.

La validez, en términos generales, se refiere al grado en que una variable representa realmente la característica a medir. Por ejemplo, si interesa medir la inteligencia y se utiliza la memoria como medida, esta no es válida.

Es válido medir la altura con una cinta métrica, pero no es válido utilizar la altura como medida de la capacidad de un aspirante a ingresar a la universidad.

Muchas veces el problema de la validez de una variable para medir un concepto se encuentra en la naturaleza misma de ese concepto, tal como ocurre con la inteligencia.

¿Qué es la inteligencia? ¿El cociente intelectual mide la inteligencia? Algunos psicólogos dirán sí. Otros argumentarán que la inteligencia está compuesta por una gran variedad de capacidades mentales y por lo tanto no puede ser medida con un único instrumento. Si no podemos decidir qué es exactamente la inteligencia, menos podremos decidir cómo medirla.

Otro ejemplo problemático es la medición del dolor. El dolor es una experiencia personal. La forma más común de medirlo es preguntarle al paciente sobre las dificultades que ese dolor le acarrea. También se le puede pedir una descripción del nivel de dolor en alguna escala, ésta no significará lo mismo para diferentes personas.

Aún así, a veces no es tan difícil hallar la variable que provee una medición válida.

La **cantidad de accidentes fatales** por año no es una variable válida si queremos evaluar los resultados de una campaña de educación vial, pues los accidentes pueden aumentar si se incrementa el parque automotor o aumenta la población, como vimos en la sección 3.1.2, la **tasa de muertes** por cada 100.000 vehículos en circulación es una medida más adecuada.

> Muchas veces una **tasa** (dada como fracción, proporción o porcentaje) es la medida válida, en contraposición con tomar simplemente **cantidades**.

□ 15.3. Más de una válida

Una variable es una medida válida de un concepto si lo representa adecuadamente o es una característica importante de dicha propiedad. Pero puede haber más de una medida válida para un mismo concepto.

Un aviso institucional televisivo anuncia “1 de cada 8 mujeres, puede padecer cáncer de mama en algún momento de su vida, pero las mujeres que tienen antecedentes familiares tienen 2 a 4 veces más riesgo”. Pero ¿cómo se mide el riesgo? Veremos dos maneras de medir el riesgo y como se calculan.

La **proporción** y la **razón** son dos medidas válidas para medir el riesgo de padecer una enfermedad.

Veamos primero cómo se calculan esas dos medidas y luego, cómo los resultados pueden ser distintos:

- Cuando se mide como una **proporción** se toma la cantidad de personas que experimentan el suceso (padecer cáncer de mama) y se lo divide por la **cantidad total** de personas en riesgo de tener el evento.
- Cuando se mide como una razón también se toma la cantidad de personas que experimentan el suceso (padecer cáncer de mama) pero en este caso se divide solamente por la **cantidad de personas que no experimentan** el suceso.

De acuerdo al aviso, el riesgo de padecer cáncer de mama para la población es:

- Cuando se mide como una **proporción:** $\frac{1}{8} = 0{,}125$

- Cuando se mide como una **razón:** $\frac{1}{8} = 0{,}143$

Una proporción de 0,125 y una razón de 0,143 son dos medidas válidas del mismo riesgo.

Siguiendo con el aviso: ¿qué significa tener 4 veces más riesgo de padecer cáncer de mama? No es lo mismo cuadruplicar la proporción que cuadruplicar la razón:

Cuadruplicar la proporción resulta en una proporción de $\frac{4}{8} = \frac{1}{2}$
La mitad de las mujeres padecerán cáncer y la otra no.

Cuadruplicar la razón resulta en una razón de $\frac{4}{7}$
4 padecerá la enfermedad y 7 no.

El aviso no aclara qué medida se utilizó.

□ 15.4. Números índices

Los **números índices** se utilizan, en forma similar a la tasa de desocupación (sección 15.1), para mostrar cómo cambia una característica con en el tiempo. Describen el cambio porcentual respecto al valor en un período base.

Tienen la ventaja de ser adimensionales, es decir, no tienen unidades. Por ejemplo, si se trata de un índice para reflejar la evolución de la superficie cubierta construida por mes, no importará si esa superficie está medida en metros cuadrados o en pies cuadrados.

Un número índice es el cociente, entre el valor de una variable en un momento del tiempo y el valor de la misma variable en otro momento llamado período base, multiplicado por 100:

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

¡Un número índice es un porcentaje!

PRECIOS PROMEDIO, MENSUALES POR LITRO DE NAFTA SÚPER EN SURTIDOR. AGO-99 A JUL-00 TABLA 15.1

Período	Precio ($/litro)
Ago-99	0,920
Sep-99	0,940
Oct-99	0,975
Nov-99	0,977
Dic-99	1,024
Ene-00	1,049
Feb-00	1,051
May-00	1,055
Jun-00	1,052
Jul-00	1,064

Secretaría de Energía. Boletín de Precios de Combustibles – junio de 2001

Veamos un ejemplo sencillo. La Secretaría de Energía releva el precio por litro de nafta súper en surtidor en 500 estaciones de servicio distribuidas por todo el país e informa mensualmente el promedio de esos precios.

El precio promedio mensual en 500 estaciones de servicio, por litro de nafta súper en surtidor, se muestra en la segunda columna de la tabla 15.1. Un litro de nafta costaba $ 0,92 en agosto de 1999 y $ 1,064 en julio de 2000. Utilizaremos los datos de la tabla 15.1 para ilustrar cómo se calculan los números índices

Tomando como período base el mes de agosto de 1999, el número índice del precio de la nafta en julio de 2000 es 115,65:

Los **números índices** se utilizan, en forma similar a la tasa de desocupación (sección 15.1), para mostrar cómo cambia una característica con en el tiempo. Describen el cambio porcentual respecto al valor en un período base.

Tienen la ventaja de ser adimensionales, es decir, no tienen unidades. Por ejemplo, si se trata de un índice para reflejar la evolución de la superficie cubierta construida por mes, no importará si esa superficie está medida en metros cuadrados o en pies cuadrados.

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

$$= \frac{1,064}{0,92} \times 100$$

$$= 115,65$$

El índice del precio de la nafta en el período base (agosto de 1999) es 100:

$$\text{número índice} = \frac{0{,}92}{0{,}92} \times 100$$

$$= 100$$

Por supuesto, los valores del índice dependen del período tomado como base.

Tomando como período base el mes de diciembre de 1999, el número índice del precio de la nafta en julio de 2000 es 103,91:

$$\text{número índice} = \frac{1{,}064}{1{,}024} \times 100$$

$$= 103{,}91$$

El **número índice** de una variable indica el valor de esa variable como **porcentaje** del valor en el **período base.**

PRECIOS PROMEDIO E ÍNDICES MENSUALES, POR LITRO DE NAFTA SÚPER EN SURTIDOR, AGOSTO DE 1999 A JULIO DE 2000 TABLA 15.2

Período	Precio ($/litro)	Índice Ago 99 = 100	Índice Dic 99 = 100
Ago-99	0,920	100,00	89,84
Sep-99	0,940	102,17	**91,80**
Oct-99	0,975	105,98	95,21
Nov-99	0,977	106,20	95,41
Dic-99	1,024	111,30	100,00
Ene-00	1,049	114,02	102,44
Feb-00	1,051	114,24	102,64
May-00	1,055	114,67	103,03
Jun-00	1,052	114,35	102,73
Jul-00	1,064	**115,65**	103,91

La tabla 15.2 muestra los precios promedio en surtidor de nafta súper junto con los números índices calculados con los datos de la tabla 15.1. En la columna 3 aparecen los índices con agosto de 1999 como período base y en la columna 4 el período base es diciembre de 1999.

El número índice 115,65 significa que el precio promedio de la nafta súper en julio de 2000 era 115,65% del valor base, es decir, el incremento respecto del valor base (agosto de 1999) es del 15,65%.

El número índice 91,80 de septiembre 1999 significa que en ese mes el valor era el 91,80% del valor base, o sea un 8,20 % menor que en el período base (diciembre de 1999).

El número índice para el período base es 100; por ejemplo si el período base es el mes agosto de 1999 se indica como "**agosto de 1999 = 100".**

Conocer el **período base** es esencial para poder interpretar un número índice.

Muchas veces se utiliza como período base un año. Por ejemplo, si el período base para un índice de precios de un litro de nafta es el año 1999 se lo indica como "1999 = 100" y el valor base es el promedio de los precios mensuales. En este caso, salvo cuando el precio promedio de algún mes coincida con el promedio anual, ningún mes tendrá índice 100.

El precio promedio de la nafta, entre las estaciones de servicio, no coincide con el gasto promedio en nafta de una persona.

15.4.1. Índice de precios al consumidor

El índice de precios al consumidor es uno de los indicadores más importantes generados por los institutos de estadísticas oficiales del mundo. Es una medida del poder de compra de la unidad monetaria, pesos en nuestro caso. Afecta las decisiones gubernamentales y está vinculado directamente con gran parte de la economía.

Inquilinos y propietarios comparan su evolución con la pactada en los contratos de alquiler, para ver quién gana y quién pierde.

Pero ... ¿Qué es el índice de precios al consumidor?

El índice de precios al consumidor es un indicador de la evolución en el tiempo, en relación a **un período base,** de los precios de la canasta familiar. La **canasta familiar** es un grupo prefijado de bienes y de servicios representativos del gasto de los hogares en una **zona de referencia.** La evolución de los precios al consumidor puede ser diferente entre provincias.

Por ejemplo, un período base puede ser el año 1999 y la zona de referencia el Gran Buenos Aires.

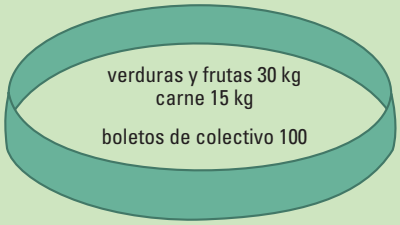
Se registran los precios de las componentes de la canasta familiar (bienes y servicios) en un período inicial o período base. Los precios de ese momento son "la base" del índice. Luego, se siguen registrando los precios a lo largo del tiempo de **la misma canasta familiar.** Para calcular el índice se comparan los precios de cada período con los precios del período base. Pero, ¿cómo se hace esa comparación?

Para contestar esa pregunta empecemos por un ejemplo. Consideremos los gastos de una familia hipotética (es decir, inventada) la familia Pérez cuya canasta familiar mensual tiene solamente 3 componentes: verduras y frutas (30 kg), Carne (15 kg), Boleto colectivo (100 viajes) (esto también es hipotético).

ESTRUCTURA DE GASTOS FAMILIA PÉREZ JULIO DE 2008 (PERÍODO BASE). TABLA 15.3

Bienes y servicios	Cantidad (julio-08)	Precio por unidad (julio-08)	Costo (julio-08)
Verduras y frutas	30 kg	$ 10/kg	$ 300 = $ (30 x 10)
Carne	15 kg	$ 20/kg	$ 300 = $ (15 x 20)
Boleto de colectivo	100 viajes	$ 1/viaje	$ 100 = $ (100 x 1)
Costo total			$ 700

La tabla muestra los costos de los bienes y servicios de la familia Pérez en el período base. En la primera columna se muestran las cantidades de cada uno de los bienes y servicios. En la segunda su precio por unidad, y en la tercera el costo total, que se obtiene multiplicando el precio unitario por la cantidad de unidades. Finalmente, el gasto total de la familia en el mes de Julio de 2008 resulta de sumar el gasto en cada uno de los bienes y servicios ($ 700).



¿Cuánto cuesta ésta canasta?

Para hallar el valor del índice en el mes de agosto de 2008, para la familia Pérez, utilizamos los precios por unidad correspondientes a ese mes (tabla 15.4, son valores inventados para ejemplificar el cálculo), con la misma cantidad de bienes y servicios de los del período base (julio de 2008).

COSTOS DEL MES DE AGOSTO DE 2008. TABLA 15.4

Bienes y servicios	Cantidad (julio-08)	Precio por unidad (agosto-08)	Costo (Agosto-08)
Verduras y frutas	30 kg	$ 11/kg	$ 330 = $ (30 x 11)
Carne	15 kg	$ 20/kg	$ 300 = $ (15 x 20)
Boleto de colectivo	100 viajes	$ 1,30/viaje	$ 130 = $ (100 x 1,3)
Costo total			$ 760

Los mismos bienes y servicios costaban $ 700 en el mes de julio de 2008 y en el mes de agosto $ 760. Por lo tanto el número índice, para la familia Pérez, agosto de 2008 (julio de 2008 =100) es 108,57:

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

$$= \frac{760}{700} \times 100$$

$$= 108,57$$

El índice mide la variación del valor de la canasta familiar respecto del período base. Para su cálculo se debe registrar el costo de la misma colección de bienes y servicios, **las mismas cantidades y los mismos productos** del período base. En el cálculo no interviene el cambio de hábitos posiblemente introducido por la familia al producirse, por ejemplo, un 30% de aumento en el costo de un viaje.

Describimos a continuación **otra forma de cálculo** del número índice. Muestra en forma explícita cómo los precios por unidad de cada producto (tablas 15.3 y 15.4) ingresan con **ponderaciones fijas:**

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

$$= \frac{330 + 300 + 130}{300 + 300 + 100} \times 100$$

$$= \frac{11 \times 30 + 20 \times 15 + 1,3 \times 100}{700} \times 100$$

$$= \frac{\frac{11}{10} \times 300 + \frac{20}{20} \times 300 + \frac{1,3}{1} \times 100}{700} \times 100$$

$$= \frac{11}{10} \times \frac{300}{700} \times 100 + \frac{20}{20} \times \frac{300}{700} \times 100 + \frac{1,3}{1} \times \frac{100}{700} \times 100$$

$$= \frac{11}{10} \times 42,86 + \frac{20}{20} \times 42,86 + \frac{1,3}{1} \times 14,29$$

$$= 108,58$$

O sea

$$\text{número índice} = \frac{\text{valor}}{\text{valor base}} \times 100$$

$$= \frac{\text{precio frutas y verduras ago08}}{\text{precio frutas y verduras jul08}} \times 42,86 + \frac{\text{precio carnes ago08}}{\text{precio carnes jul08}} \times 42,86 + \frac{\text{precio boleto de colectivo ago08}}{\text{precio boleto de colectivo jul08}} \times 14,29$$

$$= \frac{11}{10} \times 42{,}86 + \frac{20}{20} \times 42{,}86 + \frac{1{,}3}{1} \times 14{,}29$$

$$= 108{,}58$$

Expresión general para el cálculo del índice de precios, en el período t, con base en el período 0.

Con esta forma de cálculo el valor del índice de precios del mes de agosto resulta 108,58. El valor no coincide exactamente con el cálculo (1) anterior (108,57), solamente en la segunda cifra decimal, por errores de redondeo.

Expresión general para el cálculo del índice de precios, en el período t, con base en el período 0.

$$\text{número índice} = \sum_{i=1}^{n} \frac{p_t^i}{p_0^i} w^i, \quad \textbf{con} \quad w^i = \frac{p_0^i q_0^i}{\sum_{j=1}^{n} p_0^j q_0^j}$$

p_0^i = precio del producto i en el período 0
p_t^i = precio del producto i en el período t
q_0^i = cantidad del producto i en el período 0
w^i = ponderación del producto i
n = cantidad total de productos que componen la canasta

El índice mide la variación de los precios $\frac{p_t^i}{p_0^i}$ de los productos de la canasta familiar, ponderados por la participación de cada uno de los productos en el valor total de la misma en el período base.

$\left(\frac{p_0^i q_0^i}{\sum_{j=1}^{n} p_0^j q_0^j} \right)$ El símbolo $\sum$ indica la suma sobre todos los productos de la canasta (ver sección 18.1.1 para más detalles sobre el uso de $\sum$).

Las ponderaciones utilizadas en el cálculo del índice son 42,86 tanto para "verduras y frutas" como para "carne", y 14,29 para el "boleto de colectivo" (tabla 15.5). ¿Qué significan estas ponderaciones? El costo de la canasta en el período base es de $ 700 (tabla 15.3); "frutas y verduras" con un costo de $ 300 contribuye con un 42,86% del total de la canasta. Lo mismo ocurre con "carne" mientras el gasto por "boleto de colectivo" es un 14,29% del total de la canasta. ¿Por qué decimos que las ponderaciones son fijas? Porque una vez determinada la canasta, sus cantidades y sus precios en el período base (en nuestro ejemplo es julio de 2008) todos los meses se calculará el índice utilizando las mismas ponderaciones.

Cálculo de las ponderaciones de la canasta familiar. TABLA 15.5

Bienes y servicios	Cantidad (julio-08)	Precio por unidad (julio-08)	Costo (jul-08)	Ponderación (%)
Verduras y frutas	30 kg	$ 11/kg	$ 300	(300/700) x 100 = 42,86
Carne	15 kg	$ 20/kg	$ 300	(300/700) x 100 = 42,86
Boleto de colectivo	100 viajes	$1/viaje	$ 100	(100/700) x 100 = 14,29
Costo total			$ 700	100,01

Las ponderaciones de la tabla 15.5 no suman 100 debido a errores de redondeo.

Este tipo de índice, con la canasta familiar fija en sus componentes y cantidades, se denomina **índice de Laspeyres.**

Los índices de precios al consumidor en muchos países se elaboran utilizando el índice de Laspeyres.

A pesar de su nombre francés Ernst Louis Étienne Laspeyres fue un economista y estadístico alemán.

En el cálculo del índice de precios al consumidor **los bienes y servicios se mantienen fijos,** tanto en tipo como en cantidades. Estos bienes y servicios fijos son llamados **canasta familiar.**

El índice de precios al consumidor, IPC, es un número índice para el costo de un **conjunto de bienes y servicios fijo.**

¿Por qué es importante una canasta familiar fija?

Porque de esa forma la comparación es válida. Las diferencias, se deberán únicamente a la variación de los precios. Si la canasta familiar no fuera fija, no podríamos saber si un aumento en el índice se debe a un aumento de los precios, a un aumento de las cantidades consumidas, o a cambios en los productos que se consumen.

¿Cómo obtenemos una canasta familiar para representar a muchas familias? Utilizamos una canasta familiar promedio. Pero, ¡una canasta promedio representa a muchas familias y a ninguna en particular!

Veamos cómo se realizaría el cálculo de la canasta familiar para 2 familias en el mes 1.

Bienes y servicios	Familia 1	Familia 2	Promedio
Verduras y frutas	40 kg	20 kg	30 kg
Carne	0 kg	30 kg	15 kg
Boleto colectivo	80 viajes	120 viajes	100 viajes

La primera familia es vegetariana. La segunda come menos frutas y verduras, pero utiliza más viajes de colectivo que la primera. Si promediamos las cantidades para cada rubro obtendremos las cantidades que presentamos inicialmente (tablas 15.3 y 15.4).

Los institutos de estadística realizan encuestas de hogares a muchas familias para obtener una "canasta familiar promedio". Seguramente no representará a ninguna familia en particular, pero permite evaluar las modificaciones globales de los precios. Los bienes y servicios encuestados se dividen en Rubros.

En el cálculo del Índice de Precios al Consumidor (IPC), con base en el año 1999, el INDEC utilizaba los siguientes rubros:

- Alimentos y bebidas.
- Indumentaria.

- Vivienda.
- Equipamiento y mantenimiento del hogar.
- Atención médica y gastos para la salud.
- Transporte y comunicaciones.
- Esparcimiento.
- Educación.
- Bienes y Servicios Varios.

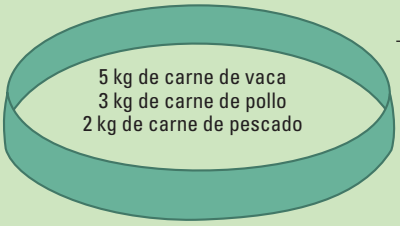
Todos ellos con una composición específica y fija.

Veamos cómo se relaciona el índice de precios, indicativo de la evolución de la economía en general, con el costo de vida individual.

Consideremos un ejemplo sencillo para ilustrar la diferencia entre el índice de precios al consumidor y el costo de vida.

El **índice de precios al consumidor** y el **costo de vida** son dos conceptos **distintos.**

Pensemos que Juan consume 10 kg de carne por mes entre carne de vaca, pollo y pescado. El precio por kg en noviembre de cada una de estas carnes es $ 25, $ 16 y $ 20 respectivamente. Por lo tanto, Juan gasta en carnes durante noviembre $ 213, distribuidos de la siguiente manera:



5 kg de carne de vaca	x $25 /kg =	$125
3 kg de carne de pollo	x $16 /kg =	$48
2 kg de carne de pescado	x $20 /kg =	$40
10 kg de carne con un gasto de		$213

Si en diciembre aumenta solamente la carne de vaca, de $25 a $30, el valor de la canasta cárnica será $238:

5 kg de carne de vaca	x 30 $/kg =	$150
3 kg de carne de pollo	x 16 $/kg =	$48
2 kg de carne de pescado	x 2 $/kg =	$40
10 kg de carne con un gasto de		$238

Por lo tanto el valor de la canasta cárnica de diciembre, para el cálculo del índice es $238.

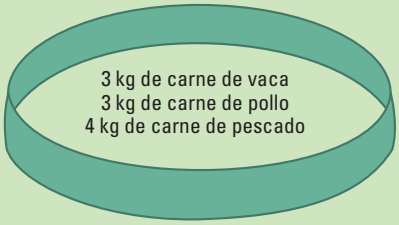
Noviembre $213 Diciembre $238

Para obtener la variación del índice se compara el valor de la canasta en el período actual con el anterior:

$$\left(\frac{238}{213}\right) \times 100 = 111{,}74$$

El aumento del IPC será del 11.74 %.

Si los precios no cambiaran Juan seguiría comprando las mismas cantidades; pero, si alguno de los precios aumenta puede **decidir cambiar.** Para un consumo total de 10 kg de carnes, podría decidir reducir su consumo de vacuna y aumentar el pescado, manteniendo su consumo total de carnes y gastar $218:



3 kg de carne de vaca x 30 $/kg = $90
3 kg de carne de pollo x 16 $/kg = $48
4 kg de carne de pescado x 20 $/kg = $80

10 kg de carnes con un gasto de $218

Juan cambió la canasta.
El índice no.

Con este gasto Juan considera que mantiene su nivel de vida y su costo de vida pasó de

Noviembre $213 Diciembre $218

En porcentaje, ese aumento es de sólo 2,3 %:

$$\left(\frac{218 - 213}{218}\right) \times 100 = 2{,}3$$

Un número índice permite resumir los valores de muchos ítems, para seguir su evolución en el tiempo. Mientras los bienes y servicios de la canasta representen adecuadamente los hábitos de la población, se mantiene fija. Con el tiempo, la canasta tiende a desactualizarse, y se requieren sucesivas adaptaciones para hacerla representativa de una realidad cambiante.

□ 15.5. Mediciones precisas y exactas

Si, por ejemplo, utilizamos la balanza de una farmacia para medir el peso dará una medida válida. Esa balanza, como ocurre a veces, puede no ser muy exacta. Si la balanza mide siempre 1 kg de más valdrá:

peso medido = peso verdadero + 1 kg

Además si repetimos la medición no obtendremos el mismo valor; la balanza no es precisa.

A veces el resultado será un poco mayor:

peso medido = peso verdadero + 1 kg + 0,25 kg

y otras un poco menor:

peso medido = peso verdadero + 1 kg - 0,75 kg

Tenemos dos tipos de errores:

Cuando las mediciones no tienen sesgo decimos que son exactas, y cuando el error aleatorio, que nunca se puede eliminar, es pequeño se trata de mediciones precisas. Ambos tipos de errores suelen estar presentes en los procesos de medición, y los podemos expresar como:

Modelo de medición: valor medido = valor verdadero + sesgo + error aleatorio

Resumiendo, un proceso de medición tiene:

- Error aleatorio, si mediciones realizadas sobre un mismo objeto dan resultados diferentes.
- Sesgo, si sistemáticamente sobreestima o subestima la propiedad que mide.

Precisión no significa validez. Por ejemplo, el volumen del cerebro no es una medida válida de la inteligencia. Sin embargo, en el siglo 19 Paul Broca sostuvo que sí lo era y propuso un método muy preciso para calcularlo (http://www.comoves.unam.mx/articulos/cerebro.shtml). Una de las consecuencias de esta idea era pensar que las mujeres son menos inteligentes que los hombres, porque sus cerebros son más pequeños (como también el resto del cuerpo), coincidiendo con los prejuicios de la época. Actualmente, no existen indicios de diferencias intelectuales entre géneros y se sabe que el volumen cerebral no tiene relación con la inteligencia.

Ningún proceso de medición es perfectamente preciso. **Los promedios mejoran la precisión.** Generalmente, los resultados de los análisis clínicos son el promedio de tres o más mediciones repetidas. Incluso, en las escuelas, los alumnos realizan varias mediciones en sus clases de laboratorio y las promedian.

Reducir el sesgo no es tan fácil porque proviene de la calidad del instrumento. En este caso, es necesario **mejorar el método de medición** para **no tener sesgo.**

¿Qué relación tiene el sesgo aquí descripto con el de la sección 6.3? ¡Se trata del mismo concepto! Dijimos en ese caso: "es un favoritismo de alguna etapa del proceso de recolección de datos". Ese favoritismo producirá una subestimación o una sobrestimación sistemática de la característica de la población a medir. La diferencia fundamental se encuentra en la interpretación del término del error. Cuando se seleccionan individuos de una población y se observa el valor de una variable, el término llamado "error aleatorio" representa las diferencias entre el valor individual (si se seleccionara sin sesgo) y la media poblacional (μ). El modelo general será:

valor individual = μ + sesgo + error aleatorio

Si los individuos se seleccionan mediante un **muestreo aleatorio** simple, eliminamos el sesgo y el modelo resulta:

$$\text{valor individual} = \mu + \text{error aleatorio}$$

Cuando se trabaja con datos es importante preguntarse: ¿Cómo se obtuvieron esos números? Si se trata de mediciones sobre muchos individuos, tendremos valores de **variables** describiendo a cada uno de ellos. Debemos saber cómo está definida exactamente cada variable y si se trata de **variables válidas** como mediciones numéricas de los conceptos en estudio.

También es necesario conocer si los datos tienen **errores de medición** que puedan reducir su utilidad. Algunos procedimientos de medición pueden introducir sesgo, en ese caso es necesario **utilizar un instrumento mejor.** Si medir al mismo individuo produce resultados diferentes, de manera que los valores no son confiables, se puede mejorar la confiabilidad **repitiendo la medición varias veces** y utilizando su promedio.

15.6 Actividades y ejercicios

1. Considerando "la inteligencia" como la capacidad de resolver problemas en general, explique por qué no es válido medir la inteligencia preguntando:

   ¿Quién escribió el Martín Fierro?
   ¿Quien ganó el último mundial de fútbol?

2. Un estudio en una ciudad muestra un promedio de 3 muertes de chicos por año en accidentes con micros colectivos y un promedio de 20 muertes en accidentes con autos particulares durante el horario escolar. Estos datos sugieren que viajar en colectivo es más seguro que viajar en auto con los padres. Sin embargo, estas cifras no cuentan toda la historia. ¿Qué comparaciones deberían hacerse para evaluar la seguridad de los dos medios de transporte?

3. El Ministerio de Salud le interesa conocer el progreso alcanzado en la lucha contra el cáncer. Algunas de las variables:

   a) Cantidad total de muertes por cáncer.
   b) Porcentaje de muertes por cáncer.
   c) Porcentaje de pacientes vivos 5 años después del diagnóstico de la enfermedad.

   Ninguna de las variables anteriores es una medida totalmente válida de la efectividad de los tratamientos. Explique cómo a) y b) podrían disminuir y c) aumentar, incluso cuando los tratamientos no fueran efectivos.

4. Interesa estudiar el "estado físico" de las alumnas de 5to año de una escuela. Describa una manera claramente inválida de medir "estado físico". Luego describa un proceso le parezca válido.

16. Variables numéricas

□ 16.1. Histogramas y distribuciones de frecuencias

La **distribución** de una variable nos dice **cuáles son los valores** que puede tomar y su **frecuencia,** es decir, cuántas veces ocurre cada uno de los valores.

Como hemos visto, las tablas de frecuencias y los gráficos (circulares, de barras) permiten conocer la distribución (ya sea en una población o en una muestra) de los valores de una variable categórica. La distribución de los valores de la variable dentro de las diferentes categorías se puede expresar en cantidades, en proporciones o en porcentajes.

Para representar gráficamente la distribución de los datos correspondientes a una **variable numérica** (discreta o continua) también se utilizan tablas de frecuencias y un gráfico similar al gráfico de barras: el histograma.

> Un **histograma** representa, en el eje horizontal, **los valores de una variable numérica** divididos en **intervalos de clase.** En forma similar a los gráficos de barras, tiene una barra sobre cada intervalo cuya **altura indica la cantidad** (frecuencia) o **proporción** (frecuencia relativa) de datos. No se deja espacio entre las barras ó rectángulos.

Cuando los valores posibles de la **variable numérica** son pocos, la altura de cada rectángulo del histograma muestra directamente la cantidad o proporción de veces que **cada uno de los valores** ocurrió. Cuando son muchos, es necesario agruparlos definiendo previamente los intervalos.

16.1.1. Variables discretas

Una variable numérica es **discreta** cuando únicamente puede tomar valores dentro de una sucesión determinada de números. La cantidad de hermanos por alumno de una escuela es una variable discreta: puede tomar los valores 0, 1, 2, 3, 4, pero nunca valores como 2,50; 7,2; 0,30.

Veremos primero un ejemplo de una variable numérica discreta (cantidad de hijos) con **pocos valores posibles. No es necesario agruparlos.**

Ejemplo 16.1. Supongamos que se entrevistan 1.000 familias de la Ciudad de Buenos Aires, para saber cuántos hijos tiene cada familia. Nuestros datos son de la forma 0, 0, 3, 1, 1, 1, 2, 2, 2, 3, 1, 1, 2, 0, 0, 0, 2, 1, 8, 1, 1, 2, 3, 0, 0, 0...

Cada número es la cantidad de hijos de cada una de las familias entrevistadas. Es necesario resumir la información: 250 familias no tienen hijos, 200 tienen 1 hijo, 300 tienen 2 hijos, 160 tienen 3 hijos, 50 tienen 4 hijos, 20 tienen 5 hijos, 10 tienen 6 hijos, 7 tienen 7 hijos, 2 familias tienen 8 hijos y una familia tiene 9 hijos. Podemos presentar el resumen mediante la siguiente tabla de frecuencias:

Cantidad de hijos	Frecuencia
0	250
1	200
2	300
3	160
4	50
5	20
6	10
7	7
8	2
9	1
Total	1.000

Tendremos una visualización más rápida de los datos si los representamos mediante un histograma.

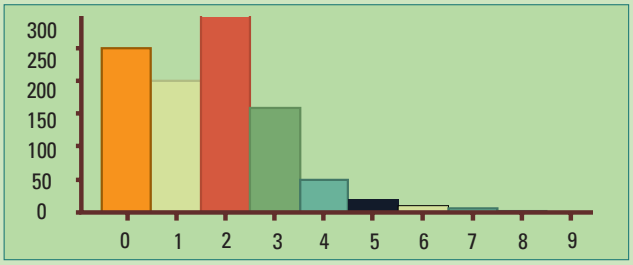


Figura 16.1. *Histograma de la cantidad de hijos por familia, expresado en frecuencias.*

La mayor cantidad de familias tienen 2 hijos, le siguen las familias sin hijos y después las de un sólo hijo.

Un histograma representa **la distribución de una variable** numérica **en una población** o **en una muestra.** Los intervalos de clase de una variable discreta están centrados en sus valores posibles y tienen la misma longitud.

En el ejemplo 16.1 los datos corresponden a una muestra de 1.000 familias de la Ciudad de Buenos Aires.

¿Cuál es la variable numérica y cuál es la población? ¿Cuáles son los valores posibles de esa variable numérica en la población? ¿Cuál es el tamaño de la muestra?:

- Variable numérica discreta: cantidad de hijos por familia.
- Valores posibles: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.
- Población: todas las familias de la Ciudad de Buenos Aires, en un año fijo.
- Tamaño de la muestra: 1.000

Si la muestra es representativa de las familias de la Ciudad de Buenos Aires en ese momento, podremos considerar al histograma, una estimación de la distribución de la variable cantidad de hijos por familia en la población. ¡Un verdadero trabalenguas!

Cuando interesa comparar la frecuencia entre categorías, como ocurre con los diagramas de barras, puede ser más interesante que el eje vertical esté expresado en frecuencias

relativas (es decir proporciones). Por ejemplo, si queremos estudiar el comportamiento social respecto a la cantidad de hijos, saber que el 75% de las familias tienen como máximo dos hijos es más informativo que saber que son 750.

Cantidad de hijos	Frecuencia	Frecuencia relativa	Porcentaje
0	250	250/1.000 = 0,250	25,0
1	200	200/1.000 = 0,200	20,0
2	300	300/1.000 = 0,300	30,0
3	160	160/1.000 = 0,160	16,0
4	50	50/1.000 = 0,050	5,0
5	20	20/1.000 = 0,020	2,0
6	10	10/1.000 = 0,010	1,0
7	7	7/1.000 = 0,007	0,7
8	2	2/1.000 = 0,002	0,2
9	1	1/1.000 = 0,001	0,1
Total	1000	1 1	100,0

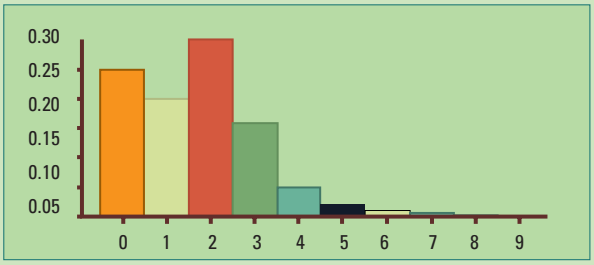


Figura 16.2. *Histograma de la cantidad de hijos por familia, expresado en frecuencias relativas.*

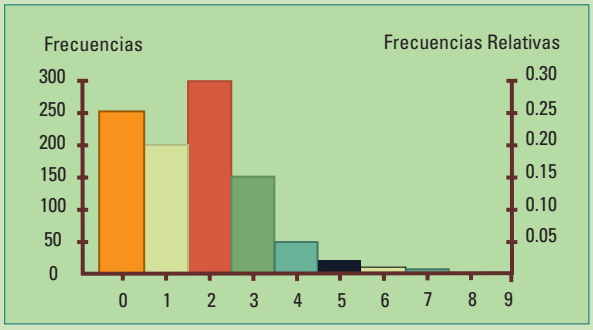


Figura 16.3. *Histograma de la cantidad de hijos por familia, con dos escalas: Frecuencias y frecuencias relativas.*

Observación. Los histogramas de frecuencias y de frecuencias relativas tienen siempre la misma forma, tal como se puede apreciar en las figuras 16.1 y 16.2. Cambian únicamente las escalas verticales. Algunas veces se presentan ambas en el mismo gráfico.

El ejemplo 16.1 (cantidad de hijos por familia) es hipotético. Como es difícil definir "familia", resulta más realista considerar la cantidad de hijos por mujer, como veremos en el siguiente ejemplo con datos reales.

Ejemplo 16.2. Se trata de la cantidad de hijos de mujeres con edades entre 30 y 34 años en el año 1991 en la Ciudad de Buenos Aires (tabla 16.1); 25.729 mujeres no tienen hijos (24,5%), 19.573 mujeres tienen un solo hijo (18,6%), 33.060 mujeres tienen 2 hijos (31,4%), etc.

El ejemplo 16.1 (cantidad de hijos por familia) es hipotético. Como es difícil definir "familia", resulta más realista considerar la cantidad de hijos por mujer, como veremos en el siguiente ejemplo con datos reales.

Ejemplo 16.2. Se trata de la cantidad de hijos de mujeres con edades entre 30 y 34 años en el año 1991 en la Ciudad de Buenos Aires (tabla 16.1); 25.729 mujeres no tienen hijos (24,5%), 19.573 mujeres tienen un solo hijo (18,6%), 33.060 mujeres tienen 2 hijos (31,4%), etc.

Cantidad de hijos de mujeres, con edades desde 30 a 34 años de la Ciudad de Buenos Aires. Año 1991. TABLA 16.1

Cantidad de hijos	Frecuencia	Frecuencia relativa	Porcentaje
0	25.729	25.729/105.210 = 0,245	24,5
1	19.573	19.573/105.210 = 0,186	18,6
2	33.060	33.060/105.210 = 0,314	31,4
3	18.020	18.020/105.210 = 0,171	17,1
4	5.467	5.467/105.210 = 0,052	5,2
5	1.867	1.867/105.210 = 0,018	1,8
6	813	813/105.210 = 0,008	0,8
7	380	380/105.210 = 0,004	0,4
8	216	216/105.210 = 0,002	0,2
9	85	85/105.210 = 0,001	0,1
Total	105.210	1 1	100,0

Fuente: *Dirección General de Estadística y Censos (G.C.B.A.) sobre la base de datos del Censo Nacional de Población y Vivienda, 1991 - Serie C.*

Un histograma de los datos de la tabla 16.1 nos permite visualizar más rápidamente su distribución.

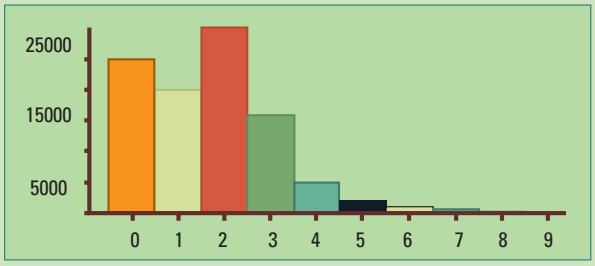


Figura 16.4. *Datos reales. Ciudad de Buenos Aires año 1991. Histograma de la cantidad de hijos por mujer con edades entre 30 y 34 años.*

La frecuencia (escala vertical del histograma, figura 16.4) es la cantidad de mujeres con edades entre 30 y 34 años en el año 1991, con 0,1, 2, ..., hasta 9 hijos, respectivamente en cada intervalo. Se destaca el rectángulo centrado en 2, porque tiene la mayor altura; 2 es la cantidad más frecuente de hijos en la Ciudad de Buenos Aires.

La distribución, es muy parecida a la del ejemplo hipotético; ambos histogramas tienen casi la misma forma pero las frecuencias, frecuencias relativas y porcentajes ya no son números redondos.

¿Cuál es la variable numérica y cuál es la población? ¿Cuáles son los valores posibles de esa variable numérica en la población?:

- Variable numérica discreta: cantidad de hijos por mujer
- Valores posibles: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 (no es posible tener 2,75 hijos).
- Población: todas las mujeres de la Ciudad de Buenos Aires entre 30 y 34 años en el 1991.

¿Puede haber mujeres con más de 9 hijos? Efectivamente, puede haber mujeres con 10 ó más hijos. En la ciudad de Buenos Aires sólo se incluye una categoría de 10 o más, porque son pocas. Para poder comparar las categorías mediante un histograma es necesario que tengan el mismo tamaño; es decir, que correspondan a la misma cantidad de valores posibles de la variable. Por esta razón no se incluyó en el histograma la categoría 10 ó más, correspondiente a los valores 10, 11, 12, 13, 14, etc.

16.1.2. Variables continuas

Una variable numérica es continua cuando, dados dos valores posibles de la variable, ésta siempre puede tomar cualquier valor intermedio. El peso de una persona es una variable numérica continua, puede tomar valores como 48 kg ó 49 kg y también, 48,5 kg 48,52 kg etc.

Podemos preguntarnos: ¿cambió la edad a la cual las mujeres tienen hijos? Veamos un ejemplo real para intentar responder esta pregunta. Como la variable edad tiene muchísimos valores posibles, para construir un histograma, los agruparemos en intervalos.

Ejemplo 16.3 Comparemos como se distribuye la edad de las mujeres en el momento del nacimiento de un hijo, en los años 2001, 2003, 2006, utilizando la información del Ministerio de Salud.

NACIMIENTOS EN LA REPÚBLICA ARGENTINA SEGÚN EDAD DE LA MADRE. TABLA 16.2

Año	2001	2003	2006	2001	2003	2006
Grupo de edad	Cantidad			Porcentaje		
[10-15)	3.022	2.763	2.766	0,44	0,40	0,40
[15-20)	97.060	92.461	103.885	14,20	13,25	14,92

Año	2001	2003	2006	2001	2003	2006
Grupo de edad	Cantidad			Porcentaje		
[20-25)	188.415	184.155	174.342	**27,57**	26,39	**25,03**
[25-30)	170.748	179.107	176.931	24,98	25,66	25,40
[30-35)	128.521	137.359	139.003	**18,80**	19,68	**19,96**
[35-40)	68.162	71.497	73.177	**9,97**	10,24	**10,51**
[40-45)	19.658	20.674	19.866	2,88	2,96	2,85
[45-50)	1.417	1.438	1.405	0,21	0,21	0,20
[50-55)	98	92	83	0,01	0,01	0,01
Sin información	6.394	8.406	4.993	0,94	1,20	0,72
Total	683.495	697.952	696.451	100,00	100,00	100,00

Fuente: *Estadísticas Vitales. Ministerio de Salud. 2001, 2003, 2006. ISSN 1668-9054.*

¿Cómo se interpretan los grupos de edad?

El grupo [10-15) corresponde a las edades entre 10 y 15 años
El grupo [15-20) corresponde a las edades entre 15 y 20 años
El grupo [20-25) corresponde a las edades entre 20 y 25 años
El grupo [25-30) corresponde a las edades entre 25 y 30 años
El grupo [30-35) corresponde a las edades entre 30 y 35 años
.....
Una edad de 15 años se cuenta en el grupo [15-20) y no en el [10-15)
Una edad de 20 años se cuenta en el grupo [20-25) y no en el [15-20)
.....

El intervalo [15-20) es un intervalo cerrado en 15 (se incluye el valor 15 en el intervalo) y abierto en 20 (no se incluye el valor 20 en el intervalo).

¿Cuál es la variable numérica y cuál es la población?:

En general, el intervalo **[a-b),** donde **a** y **b** son números reales cualesquiera con **a** menor que **b**, es un **intervalo cerrado en a** (incluye el valor **a**) y **abierto en b** (**no** incluye el valor **b**)

- Variable numérica continua: edad de la madre en el momento del parto. Es posible tener una edad decimal de 18,75 años (18 años y 9 meses).
- Valores posibles: desde 10 hasta 54 años.
- Población: se consideran en este ejemplo tres poblaciones:
- Todos los niños nacidos en el año 2006.
- Todos los niños nacidos en el año 2003.
- Todos los niños nacidos en el año 2001.

Los histogramas de la figura 16.5 permiten comparar cómo se distribuyen las edades de las madres de la República Argentina en la población de los niños nacidos en el año 2006, 2003 y 2001 respectivamente.

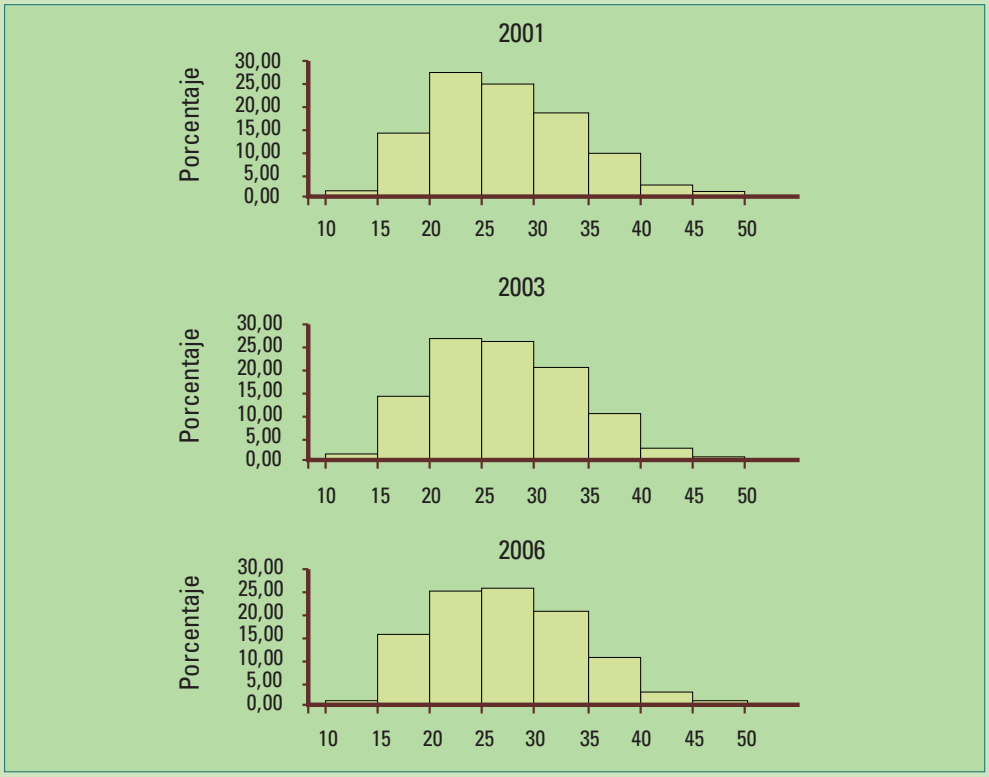
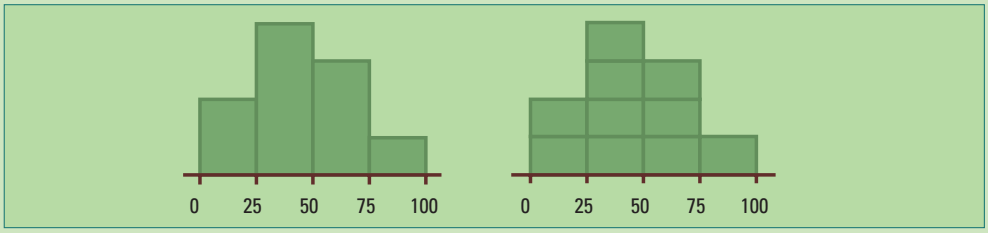


Figura 16.5. *Edad de la madre en momento del parto para los años 2001, 2003, 2006 en la Ciudad Autónoma de Buenos Aires.*

Los 3 histogramas de la figura 16.5 tienen formas similares, esto indicaría que la respuesta a la pregunta planteada es no. No cambiaron las edades en las cuales las mujeres tienen hijos en la República Argentina entre los años 2001, 2003 y 2006. Sin embargo, si observamos con más detalle vemos un porcentaje mayor en el año 2001 de nacimientos provenientes de madres con edades en el intervalo [20-25) años. En el 2003 esa diferencia entre los intervalos [20-25) y [25-30) se hace casi imperceptible y en el 2006 comienza ya el [25-30) tiene un porcentaje de 25,40 % un poco mayor que el del [20-25) con 25,03%. Mirando la tabla 16.2 (pág. 100) podemos ver además, porcentajes crecientes desde el 2001 al 2006 en los grupos de edades [30-35) y [35-40) desde el 2001 al 2006. Estas tendencias favorecen la idea que las mujeres tienen sus hijos a edades cada vez más tardías aunque se mantiene alto, cercano al 15%, el porcentaje de madres adolescentes. Esto es una preocupación de las autoridades sanitarias. La incidencia de prematuros, bajo peso al nacer y de parto instrumentado, es mayor entre las madres adolescentes que en madres con edades entre 20 y 30 años.

En un histograma puede faltar el eje vertical.

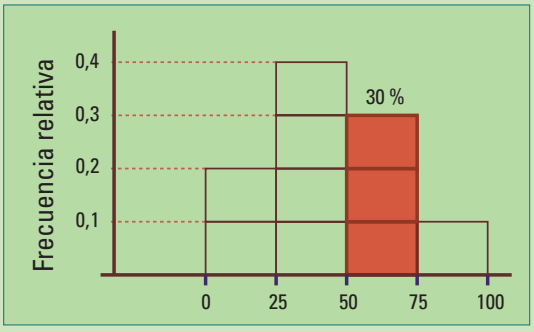
Ejemplo 16.4. Al siguiente histograma le falta el eje vertical. ¿Qué información puede proveer?



Sin el eje vertical no se pueden hallar las frecuencias absolutas, pero, sí es posible determinar las frecuencias relativas de cada uno de los intervalos. Debemos ver qué proporción del área total del histograma se encuentra por encima de cada intervalo. Dividimos la superficie del histograma en 10 rectángulos iguales de manera que cada porción es 1/10 de esa superficie, es decir el 10%.

Hay 2 rectángulos sobre el intervalo 0-25, tienen el 20% del área; 4 rectángulos sobre 25-50, 40% del área; **3 rectángulos sobre 50-75, 30%** y 10% está sobre 75-100.

Generalmente, no es tan fácil dividir a los histogramas en 10 partes iguales, sin embargo siempre las frecuencias relativas se corresponden con áreas relativas.



□ 16.2. Construcción de tablas de frecuencias

En los ejemplos anteriores los datos ya estaban agrupados o los histogramas estaban construidos. Vimos tablas con distribuciones de frecuencias para variables numéricas discretas (ejemplo 16.1 y 16.2) y para una variable numérica continua (ejemplo 16.3).

En las siguientes secciones veremos cómo se agrupan los datos numéricos y se construyen las tablas de frecuencias para obtener los histogramas. Trataremos en forma separada a los datos de variables discretas y continuas.

16.2.1. Variable discreta

Paso 1. Se ordenan los valores posibles de la variable.
Paso 2. Se cuenta cuántas veces aparece un dato con cada valor posible. Esto nos da la frecuencia.
Paso 3. Se divide cada frecuencia por el total de datos, obteniendo así la frecuencia relativa.

Ahora su turno: Registre cuántos hermanos tienen cada uno de los alumnos de su división y obtenga una tabla de frecuencias y de frecuencias relativas. ¿Cuál es la variable? ¿Cuáles son sus valores posibles? A partir de la tabla construya el histograma correspondiente. ¿Cuál es la población en estudio?

16.2.2. Variable continua

Paso 1. Se ordenan los datos.
Paso 2. Se definen intervalos de clase con igual longitud, cubriendo el rango de los valores observados.
Paso 3. Se cuentan cuantos datos pertenecen a cada uno de los intervalos. Esto indica la frecuencia.
Paso 4. Se divide cada frecuencia por el total de datos, obteniendo así la frecuencia relativa.

En el ejemplo siguiente veremos cómo construir la tabla de frecuencias para datos de una variable numérica continua.

Ejemplo 16.5. Los datos siguientes corresponden al peso (en kg) de 52 alumnos y 49 alumnas de 3 divisiones de 4to. año.

- **Varones:** 67 57 64 73 65 69 67 66 67 69 63 65 66 53 58 64 69 67 63 71 69 62 59 61 72 68 57 55 79 59 66 58 72 67 71 67 65 61 63 69 74 64 66 70 63 51 79 68 67 66 85 81

- **Mujeres:** 46 52 52 52 51 43 48 44 55 43 50 57 52 54 51 54 48 48 62 52 50 52 45 54 47 50 50 51 60 56 51 52 54 42 54 48 50 56 50 48 52 55 54 58 46 37 38 68 70

¿Cuál es la variable? Peso
¿Es una variable numérica continua o discreta? El peso es una variable numérica continua.

¿Cuál es la población?
Si nos interesa describir el peso de los/as alumnos/as de esas 3 divisiones de 4to. año, la población está formada por todos/as los alumnos/as de esas 3 divisiones.

¿Qué podemos decir de la distribución de los pesos mirando estos datos?
Para comenzar construiremos un diagrama de puntos, donde cada punto corresponde a un alumno de ese peso. Los valores repetidos se ponen uno encima del otro, a iguales distancias. ¿Se puede ver algo raro? Hay espacios vacíos y se distinguen 2 picos.

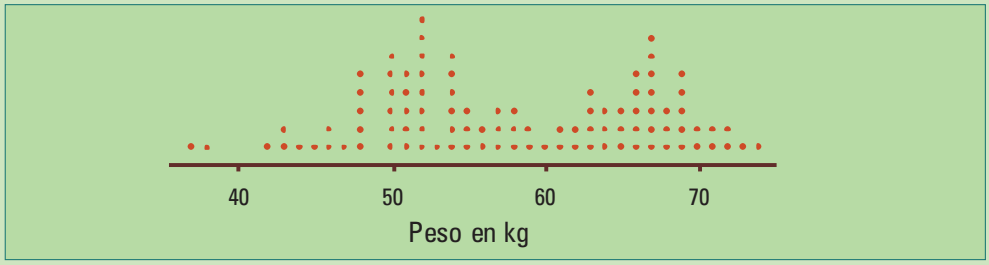


Figura16.6. *Diagrama de puntos de los pesos de varones y mujeres de 4to. año.*

Luego, construiremos una tabla de frecuencias, para eso se dividimos la recta numérica en intervalos de clase y contamos cuántos pesos caen dentro de esos intervalos. La frecuencia relativa es la proporción de pesos dentro de cada intervalo.

FRECUENCIAS DE LOS PESOS (EN kg) DE LOS ALUMNOS Y ALUMNAS DE 4TO. AÑO. TABLA 16.3

Intervalo de Clase	Frecuencia	Frecuencia relativa
[30 - 45)	6	
[45 - 60)	48	
[60 - 75)	43	
[75 - 90)	4	
		COMPLETAR
Total	101	1

El intervalo [30-45) es un intervalo cerrado en 30 (se incluye el valor 30 en el intervalo) y abierto en 45 (no se incluye el valor 45 en el intervalo).

El intervalo [45-60) es un intervalo cerrado en 45 (se incluye el valor 45 en el intervalo) y abierto en 60 (no se incluye el valor 60 en el intervalo).

El número al lado del corchete se incluye en el intervalo, el número al lado del paréntesis no.

Ahora su turno. **Completar:**

El intervalo [60-75) es un intervalo cerrado en ..... y abierto en ....., porque

.....

El intervalo [75 - 90) es un intervalo cerrado en ..... y abierto en ....., porque

.....

FRECUENCIAS DE LOS PESOS (EN kg) DE LOS ALUMNOS DE 4TO. AÑO. TABLA 16.4

Intervalo de Clase	Frecuencia	Frecuencia relativa
[30 - 35)	0	
[35 - 40)	2	
[40 - 45)	4	
[45 - 50)	9	
[50 - 55)	26	
[55 - 60)	13	
[60 - 65)	12	
[65 - 70)	23	
[70 - 75)	8	
[75 - 80)	2	
[80 - 85)	1	
[85 - 90)	1	
		COMPLETAR
Total	101	1

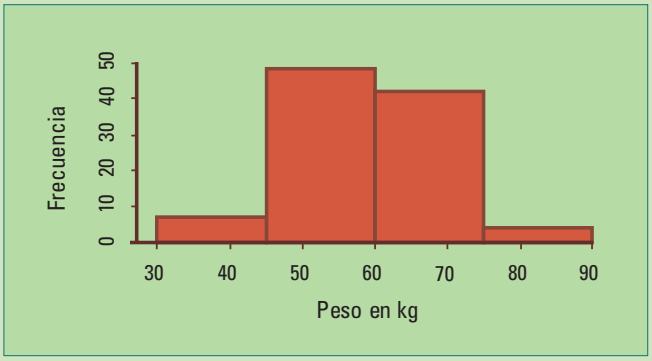


Figura 16.7.

Ahora, se debe construir el histograma. Éste (figura 16.7) no parece demasiado interesante. La mayoría de los pesos se encuentran entre los 45kg y los 75 kg, entonces podemos subdividir los intervalos de clase en tres partes iguales y obtenemos una nueva tabla de frecuencias (tabla 16.4).

El primer intervalo de clase [30-35) no tiene datos, por lo tanto ningún/a alumno/a tiene su peso dentro de ese intervalo. ¿Qué significan el corchete y el paréntesis?

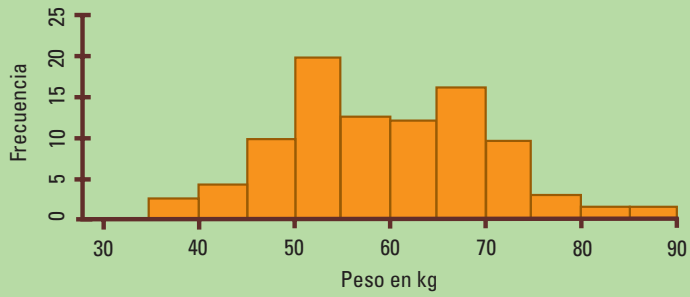


Figura16.8. *Histograma los pesos de varones y mujeres de 4to. año.*

Ahora el histograma (figura 16.8), de manera similar al diagrama de puntos (figura 16.6), nos muestra una información más interesante de la distribución de los pesos. Ambos sugieren la presencia de dos grupos aunque no se vean totalmente separados. En este ejemplo, conocemos los dos grupos mezclados, varones y mujeres. En el histograma se puede apreciar además, el carácter continuo de la variable peso.

No hay una regla para obtener la cantidad más conveniente de intervalos de clase, pero daremos unas ideas al respecto:

- Utilice intervalos de igual longitud centrados en valores redondos, si es posible, enteros.
- Si tiene pocos datos utilice una pequeña cantidad de intervalos.
- Para conjuntos de datos más grandes utilice más cantidad de intervalos.
- Una cantidad adecuada suele ser entre 6 y 12 intervalos.

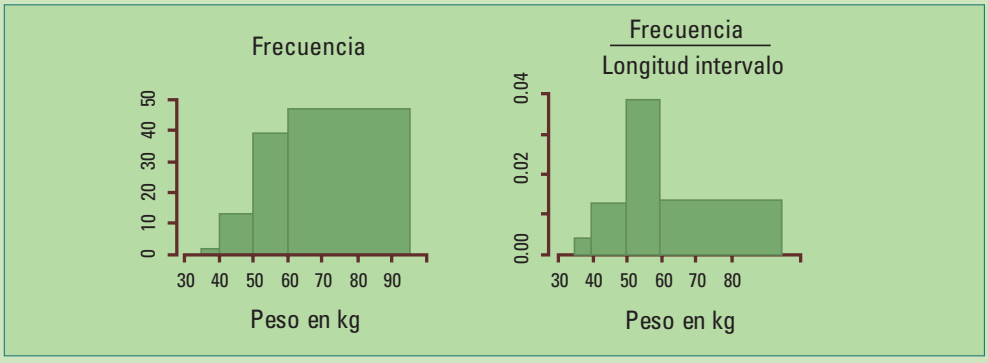
16.2.2.1. Un detalle extra

¿Pueden los **intervalos de clase** de un histograma tener **longitudes diferentes?**

Pueden, pero su construcción se complica.

En ese caso, para la altura del rectángulo de cada clase es necesario utilizar la frecuencia o la frecuencia relativa **dividida** por **la longitud** de dicho **intervalo de clase** (llamada **escala densidad**), de lo contrario, aumentar la longitud implicaría aumentar la altura, y disminuir su longitud resultaría en reducir la altura, **distorsionando** artificialmente la forma del histograma.

La figura siguiente muestra dos histogramas, en el de la izquierda la escala vertical es la frecuencia, y en el de la derecha, la frecuencia relativa dividida la longitud del intervalo de clase.



En el histograma de la izquierda, de **frecuencias absolutas** de los pesos de alumnas y alumnos de 4to. año, utilizando **intervalos de clase de distinta longitud, no representa** adecuadamente **la distribución de los datos** (ver figuras 16.7 y 16.8). Muestra más alumnos entre 60 y 90 kg que entre 30 y 60 kg. El de la derecha mejora la representación de la distribución de los datos.

Conclusión. Siempre que pueda **utilice intervalos de clase de la misma longitud.** Si no es posible elija la escala de densidad para el eje vertical.

□ 16.3. Diagrama tallo - hoja

Los histogramas son adecuados para conjuntos grandes de datos. Muestran su distribución pero se pierden los valores individuales. Para conjuntos con alrededor de 100 datos o menos, preferimos utilizar un diagrama tallo-hoja pues muestra no sólo la distribución de los datos sino también sus valores.

El estadístico John Tukey propuso en 1975, los diagramas tallo-hoja, una forma rápida para mostrar la distribución de datos correspondientes a variables numéricas, sin necesidad de obtener tablas de frecuencias, conservando todos los valores.

En estos diagramas las filas juegan el mismo papel de los rectángulos de clase en un histograma. Son como un **histograma girado 90º.** Cada fila está encabezada por un número, llamado **tallo,** a continuación se coloca una **línea vertical** y luego **las hojas.** Los valores de los tallos indican en forma compacta los intervalos de clase y tienen valores crecientes hacia abajo. Las hojas representan a los datos.

A continuación, construimos un diagrama tallo-hoja con los datos del ejemplo 16.5, el peso de alumnos y alumnas:

- **Varones:** 67 57 64 73 65 69 67 66 67 69 63 65 66 53 58 64 69 67 63 71 69 62 59 61 72 68 57 55 79 59 66 58 72 67 71 67 65 61 63 69 74 64 66 70 63 51 79 68 67 66 85 81

- **Mujeres:** 46 52 52 52 51 43 48 44 55 43 50 57 52 54 51 54 48 48 62 52 50 52 45 54 47 50 50 51 60 56 51 52 54 42 54 48 50 56 50 48 52 55 54 58 46 37 38 68 70

Intervalo	Tallo	Intervalo Tallo	Tallo
[30, 35)	3	[60, 65)	6
[35, 40)	3	[65, 70)	6
[40, 45)	4	[70, 75)	7
[45, 50)	4	[75, 80)	7
[50, 55)	5	[80, 85)	8
[55, 60)	5	[85, 90)	8

Elegimos los intervalos de clase y les asignamos su tallo

Los tallos están repetidos, aparecerán en el diagrama en dos filas consecutivas. En la fila superior van las hojas desde el cero al 4 y en la inferior las hojas desde el 5 al 9. Por ejemplo, el 5 de la fila superior representa al intervalo [50, 55] y allí se colocan las hojas (el segundo dígito) de todos los datos de ese intervalo y en la inferior se colocan las hojas de todos los datos del intervalo [55, 60].

El **tallo** es una columna de números correspondientes al primer dígito de los datos (dejamos el segundo dígito para las hojas)

Tallo los números crecen hacia abajo

3
3
4
4
5
5
6
6
7
7
8

En la segunda fila con tallo 5 se colocan 7 8 representando 57 kg 58 kg

Colocamos **el segundo dígito,** la hoja, en la fila adecuada

Tallo	Hojas
3	
3	
4	
4	
5	3
5	78
6	43
6	759767956
7	3
7	
8	

Hemos colocado los pesos de los primeros quince varones 67 57 64 73 65 69 67 66 67 69 63 65 66 53 58

Ya hemos completado el diagrama con todos los datos

Tallo	Hojas
3	
3	78
4	3432
4	688857886
5	3122210241402400112440024
5	7897598576658
6	43432113320
6	7597679569798675968768
7	3122400
7	99
8	1
8	5

Finalmente ordenamos los valores de las hojas

Tallo	Hojas
3	
3	78
4	2334
4	566788888
5	00000011111222222223444444
5	5556677788899
6	011223333444
6	55566666777777788899999
7	00112234
7	99
8	1
8	5