Matemáticas, Azar, Sociedad

Conceptos básicos de estadística

Patricia Inés Perry **Vilma María Mesa**

Felipe Fernández **Pedro Gómez**

una empresa docente®

Grupo Editorial Iberoamérica S.A. de C.V.

Bogotá, 1996

Segunda edición, junio de 1996

Matemáticas, Azar, Sociedad
Conceptos básicos de estadística
Autores: Patricia I. Perry, Vilma M. Mesa, Felipe Fernández, Pedro Gómez

Diseño carátula: una empresa docente®

Grupo Editorial Iberoamérica, S.A. de C.V.
Nebraska 199. Col. Nápoles. 03810 México, D.F.
Tel. 523-09-94
Fax: 543-11-73

una empresa docente®
Universidad de los Andes
Cra. 1 Este # 18 A - 70
Apartado Aéreo 4976 Tel. (57-1) 284-9911 ext. 2717. Fax: 284-1890
Servidor WWW: http: //ued.uniandes.edu.co
Bogotá. Colombia

ISBN

Impreso en México / *Printed in México*

Contenido

Introducción

El mundo de hoy se enfrenta, en diversos campos, a un volumen de información que cada vez va en aumento y que es necesario manejar ágil y eficientemente. La estadística, en muchos casos, se constituye en una buena opción para hacerlo. La estadística y la probabilidad son tópicos presentes en la vida diaria de los individuos. Evidentemente se hace necesario que el ciudadano común y corriente maneje apropiadamente ideas y conceptos básicos del lenguaje de la estocástica[1] y conozca y comprenda algo acerca del razonamiento estadístico. Se requiere que pueda usar las herramientas básicas de la estadística como apoyo para asimilar, criticar y contrastar la información recibida y que además las pueda aplicar en el campo del saber donde desarrollará su trabajo (Sobrino, 1994). Ya en 1981 se destacaba la importancia de incluir temas de probabilidad y estadística en los currículos escolares de matemáticas y se argumentaba que el estudio de tales tópicos proporciona aplicaciones significativas de las matemáticas en todos los niveles, proporciona métodos para trabajar con la incertidumbre, da alguna comprensión de los argumentos estadísticos que se utilizan frecuentemente y ayuda a identificar cuándo un procedimiento estadístico se ha utilizado o no de manera correcta (Shulte & Smart, 1981).

Es incuestionable la importancia de la educación estadística en todos los niveles. Sin embargo, no es tan evidente qué enseñar y cómo hacerlo de manera que los estudiantes adquieran una visión apropiada y amplia de esta rama de las matemáticas y además puedan aplicarla exitosamente al abordar problemas de la vida real en los que está presente el azar y la necesidad de razonar estadísticamente. Garfield (1995) pone de manifiesto que en muchos casos la práctica docente de profesores de estadística revela que ellos dan una importancia relativa mayor de la que reconocen a la enseñanza de algoritmos y procedimientos mecánicos y mucho menor al desarrollo de la comprensión de conceptos y del pensamiento estadístico. Tampoco se favorece el desarrollo de la capacidad para evaluar información cuantitativa ni la construcción de una visión enriquecida de la naturaleza de la estadística junto con unas actitudes positivas hacia su aprendizaje.

1 Término utilizado para referirse a la estadística y la probabilidad simultáneamente.

De acuerdo con uno de los principios del constructivismo en el aprendizaje de las matemáticas creemos que el individuo es quien construye su propio conocimiento en la interacción social que establece con otros (el profesor y sus compañeros del salón de clase). Por otro lado, hacemos eco a quienes, desde hace por lo menos una década, establecen el desarrollo de habilidades para resolver problemas como una de las metas principales del aprendizaje de las matemáticas escolares. Esos dos supuestos pretenden guiar la propuesta de diseño curricular para un curso de estadística elemental realizada por "una empresa docente" a través de su proyecto "Matemáticas, Azar, Sociedad". El libro que aquí presentamos, titulado *Matemáticas, Azar, Sociedad. Conceptos básicos de estadística,* recoge y explícita sólo algunos aspectos de esa propuesta: los temas a tratar –junto con su organización y presentación–, el tipo de problemas alrededor de los cuales se construye y consolida el conocimiento de los estudiantes y la importancia relativa de ciertos aspectos formales del conocimiento matemático para un curso elemental de estadística. Por esta razón parece necesario orientar al lector sobre algunas características de la propuesta que no están suficientemente explícitas en el libro.

Antes de entrar en detalle en aspectos de la propuesta que son interesantes para un profesor que quiera utilizar este libro, haremos una breve contextualización del mismo. Este libro de texto es uno de los resultados de un proceso de innovación curricular desarrollado en el segundo curso del ciclo de matemáticas para los estudiantes de ciencias sociales[2] de la Universidad de los Andes en Bogotá, Colombia. El proyecto "Matemáticas, Azar, Sociedad" fue concebido en 1987 por cuatro profesores del Departamento de Matemáticas de la Universidad[3] con el propósito de mejorar la enseñanza y el aprendizaje en un curso de probabilidad y estadística descriptiva. En un esquema de trabajo similar a la metodología de la investigación-acción, inicialmente, se hizo una propuesta de cambio al currículo del curso, se puso en práctica, se observaron y evaluaron los resultados y con base en la reflexión sobre ellos se hicieron modificaciones que fueron puestas a prueba y evaluadas, dando lugar a nuevos cambios. Trabajando de esa manera se realizaron varias iteraciones –no menos de ocho– hasta llegar a un diseño curricular aceptable.

A continuación se detallan los objetivos que se pretenden lograr con los estudiantes en un curso de estadística que siga de cerca esta propuesta, se hace

2 Estudiantes de antropología, ciencia política, derecho, lenguas modernas y psicología.

3 Al poco tiempo de iniciar el proyecto, tales profesores constituyeron al interior de la Facultad de Ciencias un grupo de trabajo llamado "una empresa docente", que en la actualidad es un centro de investigación en educación matemática.

una descripción breve del contenido, las metodologías y actividades que pueden ser útiles para conseguir los objetivos y el tipo de evaluación que se desprende de los demás elementos del currículo y es coherente con ellos.

Diseño curricular

Objetivos

Las teorías de cada profesor sobre cómo se aprende y cómo se debe enseñar matemáticas y en particular estadística, interactúan con las metas que plantea con respecto a la formación que debe lograr el estudiante como resultado de haber participado en el curso. De ahí la importancia de que el profesor haga explícitas las metas y los objetivos que persigue con el currículo del curso que realiza. "Matemáticas, Azar, Sociedad" considera –y en esto coincide con Garfield (1995)– que un curso de estadística elemental debe buscar resultados atinentes a por lo menos tres aspectos de la formación del estudiante, todos ellos igualmente relevantes. Estos aspectos cognitivos, sociales y motivacionales se pueden concretar en lo siguiente:

- Comprender conceptos e ideas que subyacen al empleo adecuado del razonamiento estadístico, tales como el azar y determinismo, la localización y variabilidad de información cuantitativa, la organización, resumen y descripción de datos, el papel de la distribución normal como modelo, la lógica de los procesos de inferencia y el poder real de conclusión de la estadística.
- Desarrollar la capacidad de razonar estadísticamente.
- Reconocer el papel que en la vida diaria juegan el azar, la probabilidad y la estadística y desarrollar herramientas que ayuden a manejar la incertidumbre.
- Reconocer la importancia de lograr una comunicación efectiva y expresar ese reconocimiento a través del empleo adecuado del lenguaje oral y escrito, la notación y la argumentación.
- Elaborar una visión de la estadística como herramienta útil para abordar problemas en los que interviene el azar, que incluya la convicción de que saber estadística significa aprender a comunicarse utilizando el lenguaje estadístico, resolver problemas estadísticos, obtener conclusiones y justificarlas explicando el razonamiento que las sustenta.
- Ampliar la visión acerca de la naturaleza de las matemáticas; com-

prender que ellas son más que un conjunto de verdades y algoritmos pre-establecidos.
- Aumentar la confianza en la propia capacidad para aprender cuando se estudia sistemáticamente y se trabaja en equipo.

Contenido

La propuesta curricular define un curso de estadística que involucra algunos conceptos de probabilidad. Enfatiza el significado de los conceptos y las relaciones entre ellos. Se enfoca más en la lógica que respalda a las definiciones operativas que en las deducciones formales de fórmulas y su empleo en forma mecánica y sin significado real para los estudiantes. La introducción de conceptos se inicia con una aproximación intuitiva, algunas veces empírica, hasta llegar a la institucionalización del conocimiento. Es decir, después de explicitar las ideas intuitivas de los estudiantes y de discutirlas y confrontarlas, se llega al saber aceptado como válido por la comunidad educativa. Los temas se agrupan en cinco grandes bloques.

Problemas sociales, sistemas sociales y azar. Con estos temas se busca hacer una reflexión sobre la naturaleza de los problemas sociales y la necesidad de simplificar la complejidad que los caracteriza. Surge el concepto de sistema social como un modelo posible que identifica los elementos relevantes y las relaciones más importantes del problema que se está simplificando. En conexión con este tema, se aborda el concepto de azar, su naturaleza, su presencia en la vida cotidiana y el papel que juega en ella. Para el estudio del azar, se puede recurrir a lecturas tales como: "Ateo, gracias a Dios" de Luis Buñuel y "Un diálogo azaroso", versión adaptada de un texto de Henri Poincaré. Esas dos lecturas se encuentran en Perry et al. (1990). Se recomiendan otras lecturas como por ejemplo, "El orden nació del caos" de Ilya Prigoguin y "La ciencia está atascada desde hace veinticinco años" de Rene Thom; ambas son capítulos de Guy Sorman (1991). Otro libro del que se pueden obtener ejemplos interesantes sobre el azar es Al Azar. La suerte, la ciencia y el mundo de Ivar Ekeland (1992).

Población, muestra y variable. Son conceptos que, por estar presentes de manera explícita o implícita en todo lo que se hace en estadística, reciben una gran atención a lo largo de todo el curso. Ellos se retoman permanentemente. Aunque parecen ser conceptos fáciles de comprender y de manejar, los estudiantes tienen dificultad para identificarlos en situaciones problemáticas concretas.

Conceptos de estadística descriptiva y probabilidad. Se inicia el estudio de

estos temas con una aproximación intuitiva y en lo posible, empírica. Para ello se realizan alrededor de ocho talleres[4] en los que los estudiantes, trabajando en grupos pequeños, se enfrentan a problemas nuevos que deben resolver. Después del trabajo en grupos se hacen puestas en común con el fin de socializar el trabajo y de ir elaborando o modificando las ideas intuitivas de los alumnos. Sin embargo, aún no es el momento de la formalización. La formalización e institucionalización del conocimiento se hace posteriormente con base en las lecturas de los correspondientes capítulos. Se incluyen temas de organización y resumen de datos, el concepto de probabilidad y sus propiedades más importantes[5] y la descripción de conjuntos de datos cuantitativos a través de medidas de tendencia central y de dispersión.

Distribución normal. El énfasis está puesto en lo que significa lograr un modelo para representar la tendencia del comportamiento de ciertas variables y en la utilidad que tiene el modelo normal estándar. Se quiere que el estudiante maneje los procesos de estandarización y desestandarización sin necesidad de recurrir a la aplicación mecánica y sin sentido de una fórmula. Es conveniente iniciar la comprensión de esos dos procesos cuando se estudian la dispersión y la desviación estándar. Establecer la relación entre aquéllos y la desviación estándar contribuye a dar significado real al concepto de desviación estándar.

Introducción a la inferencia estadística. Se quiere que el estudiante vea y reconozca en la inferencia estadística una forma de razonar que hasta ahora no ha utilizado en los cursos previos de matemáticas. De manera empírica, se induce el concepto de distribución muestral y se presenta el significado del teorema del límite central. Además se estudia la lógica que subyace a las pruebas de hipótesis y a los intervalos de confianza y se hace inferencia sobre la media y la diferencia de medias empleando el modelo normal.

Metodología y actividades

Una vez definidos los objetivos que se pretenden lograr con el curso en la formación de los estudiantes y establecido el supuesto muy general acerca de cómo se aprende (el sujeto construye su propio conocimiento en la interacción social con otros) cabe preguntarse acerca de cómo debe ser la enseñanza para

4 En el capítulo "Situaciones problemáticas" (último de este libro) se incluyen algunos de estos talleres.

5 Para este tema se utiliza lo hecho en Perry *et al*. (1990). *Matemáticas, Azar, Sociedad. Una introducción empírica a los conceptos de probabilidad*. (pp. 267-276). Bogotá: "una empresa docente".

asegurar el logro de tales objetivos de manera coherente con la naturaleza del aprendizaje. Creer que esa pregunta tiene una sola respuesta y que por tanto existe una única manera óptima de enseñar supone reducir el problema de la enseñanza y el aprendizaje a términos muy simples cuando esos procesos son fenómenos que implican al ser humano como ser social inmerso en una cultura y que por esa razón son muy complejos. Sin embargo, se pueden establecer pautas de enseñanza que insinúan dar buenos resultados de aprendizaje (Burril, 1990). Se pueden agrupar en dos clases: las que se refieren a aspectos propiamente didácticos y las que se refieren a aspectos de interacción en el salón de clase.

Aspectos didácticos

Papel del profesor. El papel del profesor se debe centrar sobre todo en la construcción de situaciones de aprendizaje (situaciones didácticas) suficientemente ricas y diversas que permitan el surgimiento de las ideas intuitivas del estudiante y el enriquecimiento de su comprensión.

Naturaleza de las situaciones didácticas. Las situaciones didácticas planteadas al alumno deben tener en cuenta sus intereses y asuntos que tengan significado para él. Deben dar al estudiante la oportunidad de experimentar previamente y trabajar con técnicas sencillas de conteo y tabulación de datos y de construcción de gráficas. Deben dar una visión amplia, no cerrada, de la naturaleza de las matemáticas y, en particular, de la estadística en la medida en que las soluciones no se obtengan de manera mecánica al aplicar algoritmos establecidos previamente en la clase. Deben dar la oportunidad de trabajar individualmente pero también –y esto es muy importante– de hacerlo en pequeños grupos de manera que sea posible discutir y confrontar las propias ideas con las de los compañeros. También se requiere abrir espacios para la comunicación y la exposición de argumentos tanto en forma oral como en forma escrita.

Realización de un proyecto de investigación. Paralelamente al desarrollo del contenido del curso, los estudiantes deben realizar un proyecto de investigación. Con respecto a este punto la propuesta "Matemáticas, Azar, Sociedad" se ha implementado como se describe a continuación. Se escoge un problema[6] –relacionado con el ambiente universitario– que se pueda abordar desde la estadística con los conceptos que se tratan en el curso. Se forman grupos interdisciplinarios de cuatro o cinco estudiantes. Ellos deben hacer un trabajo individual y también un trabajo en equipo. Para apoyar el desarrollo del proyecto se dedican algunas horas de clase para explicar y acordar cuestiones generales, para hacer el diseño metodológico y coordinar el trabajo de los diferentes grupos. Inicialmente se hace un trabajo para la identificación y definición del

problema y de los objetivos que puede tener el estudio. Luego, se dedica tiempo al diseño de la investigación. Después, entre todos los estudiantes del curso (más o menos 170 distribuidos en seis secciones) se recoge la información necesaria. Cada grupo usa la muestra tomada por todos los estudiantes para adelantar la investigación particular que se haya planteado. Para terminar, cada grupo hace una exposición de su estudio ante sus compañeros y entrega un informe final[7]. El profesor coordina el trabajo de los diferentes grupos.

Manejo de las situaciones didácticas por parte del profesor. El énfasis en cualquier trabajo de estadística debe recaer en el análisis y la comunicación del mismo y no en simples respuestas. En tanto sea posible deben utilizarse datos reales para los trabajos en estadística. Las experiencias de clase deben ser tales que ayuden al estudiante a aumentar su autoestima y confianza en su propia capacidad para aprender. El manejo del error, la formación de hábitos de estudio, el manejo de la interacción en el salón de clase son elementos que influyen significativamente en la motivación y en las actitudes de los estudiantes hacia la estadística y su aprendizaje.

Aspectos de interacción

Discusión en clase. Esta es la metodología predominante. Para participar en las discusiones que se llevan a cabo en la clase, cuando se introducen temas nuevos, cada estudiante debe preparar el tema con anterioridad leyendo lo que se estipula y respondiendo la guía de lectura propuesta. Se pretende que ese trabajo de preparación dé al estudiante la posibilidad de hacer una reflexión sobre temas aún no tratados en clase, le lleve a explicitar cuáles son

6 En los cursos que se han desarrollado siguiendo la propuesta curricular que aquí se presenta se han trabajado, entre otros, los siguientes problemas: 1) Existe polémica nacional alrededor de la dosis personal de droga. Se quiere conocer la aceptación de esta medida en la población uniandina y cómo se relaciona esta opinión con el uso de la droga, el cigarrillo y el alcohol. 2) Los estudiantes que traen vehículo a la Universidad enfrentan problemas originados por la falta de espacio para el estacionamiento, los altos costos y la inseguridad. Se quiere examinar qué tan real es el problema. 3) La Oficina X de la Universidad está interesada en determinar si las actividades en las que se involucran los estudiantes por el hecho de estar en la Universidad son situaciones que generan estrés en ellos. 4) ¿Están informados los estudiantes uniandinos de lo que ocurre en el país? Un grupo de profesores de la Universidad encomienda a los estudiantes del curso de estadística, realizar un estudio con el objetivo principal de describir la situación de interés.

7 Se pide a los estudiantes que tanto para el informe final como para el levantamiento de los datos y su procesamiento utilicen el computador.

sus pre-conceptos y además le lleve a identificar posibles dudas y preguntas. Usualmente el profesor es quien coordina la discusión y hace la institucionalización del conocimiento.

Trabajo en equipo durante la clase. Es la metodología propuesta para los talleres. Se hacen grupos de dos o tres estudiantes con la tarea de buscar entre ellos solución a algún problema o a preguntas específicas. Lo que se pretende es que entre ellos se genere una discusión y se llegue a una respuesta de grupo. Posteriormente, se lleva a cabo una discusión plenaria sobre las soluciones o se hace exposición de las mismas. El papel del profesor consiste en monitorear los procesos de discusión en los grupos pequeños para detectar su desarrollo y poder tomar decisiones con relación a aspectos que requieran mayor discusión dentro de los grupos o ampliación al grupo total o que deban considerarse en la institucionalización. También debe detectar dificultades en la apropiación de conceptos, en el uso de estrategias y en las relaciones sociales de los grupos y entre ellos. Los trabajos que se formulan tienen fundamentalmente dos propósitos: identificar dificultades de comprensión o manejo de conceptos por medio de presentación de situaciones problemáticas diseñadas con tal fin, y consolidar la comprensión de un cierto tema.

Evaluación

El diseño curricular propuesto enfatiza la coherencia entre la evaluación y los demás elementos del currículo a nivel del salón de clase, a saber, los objetivos, la metodología y el contenido.

Si se proponen objetivos cognitivos, sociales y motivacionales y además, los primeros no se centran en repetir lo que el profesor dice en clase o lo que los autores del texto afirman, es necesario que la evaluación no se limite a considerar aspectos cognitivos y más exactamente a detectar qué tanto recuerdan los estudiantes los procedimientos y los conceptos definidos en clase. Entonces, ¿qué otros aspectos podrían considerarse en la evaluación? Hay muchos; a manera de ejemplo, citaremos tres. Dado que la metodología propuesta sugiere que el estudiante haga una reflexión que le permita explicitar sus ideas intuitivas sobre un tema antes de que se trate en clase, la evaluación de ese trabajo[8], no necesariamente tiene que centrarse en qué tan correctas son

8 La forma de concretar esa reflexión en algo escrito puede hacerse de diversas maneras. Una, es a través de un resumen de la lectura donde se expongan las ideas centrales y las dudas que surgieron al lector. Otra forma es responder por escrito sólo aquellas preguntas de la guía de lectura cuyas soluciones no sean evidentes. Una tercera es hacer un mapa conceptual.

las respuestas. En cambio, puede ser interesante ver qué tan coherente y completo es el discurso, qué tipo de argumentos se dan, qué recursos se utilizan para explicar las afirmaciones, etc. También resulta interesante evaluar, ya no el trabajo que ellos hicieron como tarea, sino las correcciones, los apuntes y comentarios que hacen sobre su tarea escrita como resultado de haber participado en una discusión plenaria. Al evaluar la exposición de un alumno o un grupo de alumnos ante el resto de sus compañeros se puede evaluar algo más que el contenido de la presentación; se pueden considerar puntos tales como los recursos utilizados en la preparación y en la presentación misma, la estructura y el desarrollo de la presentación, la creatividad, etc.

En realidad, la propuesta curricular "Matemáticas, Azar, Sociedad" impulsa la evaluación en dos sentidos. Por un lado, se pretende que refleje el trabajo y el compromiso del estudiante con su aprendizaje en el curso; y por otro, que dé información –tanto al profesor como al estudiante– acerca de cómo se están desarrollando los procesos de enseñanza y aprendizaje para detectar oportunamente dificultades y logros y así poder tomar decisiones pertinentes. Por tanto, para que la evaluación cumpla esas dos funciones primordiales, se sugiere buscar diversas fuentes entre las que se pueden mencionar: tareas diarias, pruebas escritas, talleres, proyecto de investigación, exámenes, exposiciones, aspectos de la interacción social dentro del salón de clase y de las actitudes de los estudiantes ante su aprendizaje y ante el curso del que hacen parte.

Al ser un canal de comunicación entre el profesor y los estudiantes, la evaluación se convierte en un medio importante para la construcción del contrato didáctico. A través de la evaluación el profesor le indica a los estudiantes que es lo que él considera relevante en el aprendizaje de la estadística. Por su parte, el estudiante puede utilizar la evaluación para informar al profesor acerca de sus intereses, sus capacidades y sus dificultades.

Conclusiones

Esperamos que el conocimiento de algunos aspectos curriculares que subyacen a este libro dé ideas y pautas de manejo al profesor que quiera utilizarlo en un curso y también le ayude al estudiante que lo sigue a comprender por qué y para qué este libro se sale de los esquemas tradicionales de presentación a los que está acostumbrado.

Este libro es tan sólo uno de los resultados de un proyecto de innovación curricular; está estrechamente ligado con el diseño de un curso –descrito breve y parcialmente en los párrafos anteriores– y de ninguna manera pensamos que sea auto-suficiente. Los autores somos conscientes de las deficiencias que tiene; por ejemplo, hace muy poca referencia al diseño de experimentos y por eso, aunque no enfatiza en fórmulas y en cambio destaca el significado y sentido de los conceptos y de las relaciones entre ellos, el libro es, en todo caso, un compendio de herramientas y la estadística **no** es sólo herramientas. Sin embargo, la utilidad que puede prestar depende en gran medida de la forma en que el profesor lo maneje, pues él es quien interpreta, adapta y modifica las propuestas hechas.

Al final del libro hemos puesto unas referencias bibliográficas con el propósito de sugerir la consulta de artículos y libros sobre la educación estadística.

Para terminar, los autores queremos reconocer la colaboración que nos han prestado los profesores de la Universidad de los Andes que han realizado el curso con el diseño curricular propuesto. A su interés y contribución debemos los cambios que han enriquecido el diseño; ellos son coautores del capítulo "Situaciones problemáticas". Queremos agradecer específicamente a Luisa Andrade, Claudia Arévalo, Cecilia Corvalán, Martha Espinosa, Camilo Gutiérrez, Alejandro Mateus, Cesar Muñoz, Claudia Rebolledo y David Ricaurte. Agradecemos también a los estudiantes que han prestado su valiosa ayuda como monitores de nuestros cursos y a los que han participado en la realización de pequeños proyectos para la formulación de problemas de interés. Por supuesto, damos las gracias a todos los estudiantes que han tomado el curso siguiendo la propuesta, pues ellos son quienes han dado la realimentación necesaria.

Los autores

Bogotá, enero de 1996

Los problemas sociales

Introducción

Este capítulo consta de una breve lectura y de un conjunto de preguntas relacionadas con ella. Se pretende poner de manifiesto la existencia de cierto tipo de problemas que son de interés para los investigadores en ciencias sociales y se quiere delimitar de alguna manera las características de tal tipo de problemas, estableciendo simultáneamente diferencia entre ellos y los problemas que son de interés para las ciencias naturales. Además, se entreve ya la necesidad de lograr un tratamiento especial para los problemas sociales.

A ciencia cierta

(Stadi Shka y Ana Liza se encuentran en un pasillo, a la salida de clase.)

Stadi Shka: ¡Hola!, Ana Liza, acabo de salir de clase de proceso político y antes estuve en clase de física. Y, estoy en una confundida terrible...

Ana Liza: ¿Y eso, por qué Stadi Shka?

Stadi Shka: En ambas clases hablan de *problemas* que hay que resolver, pero no veo con claridad la relación entre ellos. Además, en física hablan de *ciencia*, y también lo hacen en proceso político. En este curso hablan de la *ciencia social*, mientras que en física hablan de *ciencia natural*. Pero, me parece que no tienen ninguna relación esos dos cursos. Tú, ¿qué piensas de esto?

Ana Liza: Pues, creo que tanto la física como la ciencia política son ciencias. Sin embargo, son ciencias diferentes, particularmente porque tratan problemas diferentes. Como nos interesan las ciencias sociales, te propongo que tratemos de analizar los problemas que estas ciencias abordan. ¿Qué se te ocurre?

Stadi Shka: Lo primero que se me ocurre es que los problemas de las ciencias sociales son mucho más **complejos** que los problemas de las ciencias naturales. Mira, te doy un ejemplo: en clase de física estamos viendo la caída libre de los cuerpos. A mí me parece sencillo. Para este caso, conocemos las leyes que rigen el proceso, hay muy pocas variables involucradas (la fuerza de gravedad, la masa del cuerpo en cuestión, la altura), las leyes han sido plenamente corroboradas, podemos experimentar con el proceso tantas veces como se quiera y, por consiguiente, nos es posible *predecir con certeza*. En proceso político estamos analizando la elección popular de alcaldes. En este caso, no conocemos leyes que rijan el proceso, hay muchas variables involucradas (situación económica del municipio, situación geográfica del mismo, grado de urbanismo, aspectos sociales, aspectos políticos, y muchas otras), las interrelaciones entre estas variables son innumerables, no nos es posible experimentar y, por consiguiente, resulta prácticamente imposible predecir con exactitud. Por todo lo anterior, digo que la clase de proceso político me parece complicadísima.

Ana Liza: Pues sí; los problemas de las ciencias sociales son más complejos que los de las ciencias naturales. ♣*¡Sin exagerar, niñas!, porque en niveles avanzados de las ciencias naturales hay problemas tanto o más complejos que los de las otras ciencias*.♣ Pero, ¿de dónde sale esa complejidad?

Stadi Shka: Para comenzar, hay que tener en cuenta que en los problemas sociales interviene **el hombre** y el hombre es muy complejo. O, ¿es que no has visto cómo se comporta Askanio?

Ana Liza: Pero, ¿tú crees que la complejidad de los problemas venga exclusivamente de que en ellos interviene el hombre?

Stadi Shka: No; la verdad es que la cosa es más complicada aún. Porque no interviene sólo un hombre, sino que intervienen **muchos hombres**.

Ana Liza: Sí. Pero en física hay problemas en que intervienen muchos átomos. Y la cosa no es tan complicada.

Stadi Shka: En física todos los átomos son iguales o se comportan de manera muy similar. En tanto que los hombres son diferentes unos de otros, tienen criterios y formas de pensar diferentes; más aún, interactúan entre ellos de maneras diferentes. ¡Eso es lo que hace que los problemas sociales sean tan difíciles de manejar! ¿Te das cuenta Ana Liza?

Ana Liza: Sí, claro que me doy cuenta. Evidentemente los problemas de las

ciencias sociales son muy diferentes de los problemas de las ciencias naturales. Para resumir, en los problemas de las ciencias sociales interviene el hombre, el hombre es complicado por naturaleza, normalmente intervienen muchos hombres, todos son diferentes y, para acabar, las relaciones entre los hombres suelen ser muy complejas. Qué problema, ¿no? Pero ¿qué significa esto Stadi Shka? ¿Quiere decir que mejor deberíamos estudiar física o química y dejar a un lado la ciencia política?

Stadi Shka: ¡Uy, no! Ni de riesgos. Fíjate que al estudiar ciencia política no se nos ensucian las uñas en los laboratorios... Mentiras, ese era un chiste malo. Sin embargo, me da la impresión de que debemos tratar de abordar los problemas de las ciencias sociales de manera diferente a como se hace con los problemas en las ciencias naturales. Por ejemplo, el otro día, en un programa de televisión explicaron cómo el color de las alas de una mosca —que ya no me acuerdo cómo se llama— depende de la posición de unas moléculas en no sé qué sitio de no sé qué células. ¿Tú crees que eso se puede hacer en ciencias sociales? ¿Crees que podamos explicar por qué eligieron a algún alcalde a partir de la personalidad y los gustos de cada uno de los habitantes del municipio en cuestión?

Ana Liza: Claramente no. Creo que has dado con la clave del asunto. En las ciencias sociales, los problemas que ellas tratan no se pueden **reducir** a sus últimas consecuencias. Es muy difícil **meterse por dentro** de los problemas. Me da la impresión de que lo mejor es **mirarlos desde afuera** y tratar de **simplificarlos** lo más posible para llegar a definir algo que llaman un *sistema social*. Por ejemplo, en el caso del color de las alas de una mosca, podemos llegar a conocer la composición molecular de las células de las alas. A partir de esta composición molecular y del conocimiento que se tiene del comportamiento de los átomos cuando la luz incide en ellos, podemos *deducir* cuál debería ser el color que percibimos cuando observamos las alas de la mosca. En este caso, nos hemos introducido al interior del sistema y hemos sido capaces de *predecir* a partir del conocimiento que tenemos del comportamiento de cada uno de los elementos y cada una de las interrelaciones que intervienen en el mismo. En el caso del problema de la elección popular de alcaldes, esto no sería posible. Por ejemplo, tendríamos que conocer el estado, el comportamiento y las interrelaciones de cada una de las neuronas de cada uno de los electores. Es por ello que, en este caso, nos vemos obligados a analizar el problema desde afuera con el propósito de simplificarlo al construir un sistema social.

Stadi Shka: Mira, allá viene Chepa. Veamos qué tanto sabe ella sobre sistemas sociales.

a. Explique qué quiere decir que un problema social sea complejo. Además, proponga un problema social que sea de su interés, diferente al que se menciona en el diálogo, y diga en qué consiste la complejidad de tal problema.

b. Explique por qué un problema social es complejo.

c. Establezca todas las diferencias que pueda entre los problemas de las ciencias sociales y los de las ciencias naturales.

d. Haga un comentario crítico sobre la comparación que se hace en el diálogo con respecto a la complejidad de los problemas que son de interés para las ciencias naturales y para las sociales.

e. ¿Qué cree usted que significa la expresión "mirar el problema desde afuera"?

Los sistemas sociales

Introducción

Como consecuencia de la complejidad de los problemas sociales se hace necesario buscar una manera adecuada de abordarlos. En este capítulo se presenta la construcción de un modelo —sistema social— como la mejor forma de aproximación al análisis de un problema social. Además, se presenta la estadística como una herramienta para manejar conjuntos de datos, obtener generalizaciones y sustentar numéricamente las conclusiones inferidas a partir de ellos. Este capítulo, al igual que el anterior *Los problemas sociales* y el siguiente *Ejemplos de sistemas sociales* debe considerarse como un primer intento para lograr una caracterización de los problemas sociales cuya solución se puede realizar empleando la estadística.

Del problema social al sistema social

(*Chepa se une a la conversación de Stadi Shka y Ana Liza.*)

Stadi Shka: ¡Hola Chepa! Apuesto a que tú tienes una idea clara acerca de qué es un sistema social y cómo se construye.

Chepa: Pues, de verdad, no recuerdo bien toda esa historia. Precisamente tenía la esperanza de que ustedes me explicaran pues tengo una evaluación sobre ese tema pasado mañana y me voy a rajar.

Ana Liza: No te preocupes Chepa. Creo que, aunque no nos acordemos bien de qué se trata, a partir de lo que hemos hablado con Stadi Shka, podemos descubrir muchas cosas acerca de los sistemas sociales. Para comenzar, hace un rato concluimos con Stadi Shka que una de las características de un problema social es su complejidad y que esa complejidad no nos permite, al con-

trario de lo que sucede con los problemas de la ciencia natural, mirar el problema por dentro, esto es, conocer todas las causas y el funcionamiento interno del fenómeno que queremos estudiar.

Stadi Shka: Concluimos entonces que es necesario hacer por lo menos dos cosas: mirar el problema desde afuera y tratar de simplificarlo. Para mirar el problema desde afuera tendremos que limitarnos a determinar los *elementos* y las *interrelaciones* relevantes, y además creo que en la parte de simplificación del problema interviene el concepto de sistema social. A mí esto me recuerda toda una historia en la que insistió mi profesor de matemáticas acerca de *modelar* situaciones complejas.

Ana Liza: Es verdad. Siempre que se tenía una situación compleja, para llegar a conocerla, a manejarla, a analizarla y a solucionarla, lo primero que nos aconsejaba era que tratáramos de simplificarla a través de un *modelo*.

Chepa: Sí. Entonces podemos considerar que un sistema social es un modelo de un problema social. El problema es que no me acuerdo de la historia de los modelos.

Ana Liza: Bueno, no importa. Veamos cómo se puede caracterizar un sistema social. Tal vez, lo mejor sea considerar un ejemplo. ¿Qué tal si pensamos en nuestro curso de matemáticas y en las notas del parcial que vamos a tener la semana entrante?

Chepa: Pues a menos que ustedes me expliquen, yo tengo una idea muy clara de cuál va a ser mi nota en ese parcial.

Stadi Shka: No te preocupes, Chepa. Vas a ver que con Ana Liza, uno entiende rápido. Pero, pensemos: ¿cuáles son las características de nuestro sistema social?

Chepa: Eso es fácil. En primer lugar, hay unos elementos involucrados en el sistema social: por ejemplo, los estudiantes, el profesor, los temas del curso, las calificaciones. Y, debe haber más, pero por ahora no se me ocurren.

Stadi Shka: Y, en segundo lugar, hay una cierta cantidad de interrelaciones entre esos elementos. Por ejemplo, la importancia que cada profesor da a cada uno de los temas, el tipo de preguntas que cada profesor tiene la costumbre de hacer, los objetivos que cada profesor ha definido para el curso y varias otras.

Ana Liza: De acuerdo. Cada sistema social tiene elementos y entre ellos hay unas interrelaciones. Sin embargo, como el sistema social es un modelo simplificado de un problema social es necesario determinar cuáles son los elementos y las interrelaciones que se consideran realmente **relevantes** para el problema social que se está modelando.

Chepa: Pero, falta considerar una parte vital de todo sistema social. Veamos el ejemplo de las notas del parcial: si el día del parcial llueve sorpresivamente, alrededor de la hora de clase, puede haber estudiantes que llegan tarde, no alcanzan a responder todo el parcial y por tanto, la calificación general del parcial se vería afectada por un *factor externo*.

Stadi Shka: O, también es posible que para el día del parcial tengamos que entregar varios trabajos para otros cursos.

Chepa: Podemos entonces concluir que para definir un sistema social debemos construir un modelo en el cual se determinen los elementos y las interrelaciones entre ellos y además los factores externos que influyen sobre aquéllos. Y, lo que nos interesa es analizar **cómo se comportan los elementos y las interrelaciones entre ellos dadas unas influencias externas.**

Ana Liza: Para poner otro ejemplo de factores que influyen en el sistema social de nuestro ejemplo, podríamos pensar en que el profesor decida hacer un parcial conceptual o práctico. Es posible que las notas del parcial dependan de estos factores y que al profesor le interese saber qué tipo de parcial hacer. Yo me pregunto si el profesor puede predecir cuál va a ser la nota del parcial, si hace un parcial conceptual.

Stadi Shka: Pues claro que no, Ana Liza. En los sistemas sociales, por su complejidad, es imposible predecir con absoluta certidumbre. ¡Fíjense!, esa es otra diferencia con los problemas de la ciencia natural. Si se lanza una piedra hacia arriba, con toda seguridad se sabe que después de un tiempo, ésta caerá. En cambio,...

Chepa: Bueno, sí. Pero si no se puede predecir con certidumbre, entonces ¿para qué sirve todo lo que hemos descubierto acerca de los problemas sociales?

Stadi Shka: Pues sirve y mucho, porque predecir con absoluta certidumbre no es la única manera de predecir. También se puede *predecir con algún grado de certidumbre*, con alguna probabilidad de equivocarnos. El profesor puede, por

ejemplo, predecir con una probabilidad del 80%, que si hace un parcial práctico, una tercera parte del curso se raja.

Ana Liza: Y que si hace un parcial conceptual, dos terceras partes se van a rajar con un 80% de probabilidad.

Chepa: Pero, todavía no entiendo por qué no se puede predecir con absoluta certidumbre.

Ana Liza: La razón es sencilla. El problema social es tan complejo que en una gran medida se ve influido por el *azar*. Por consiguiente, si se analiza un sistema social, correspondiente a un determinado problema social y con base en ese análisis se quiere hacer una predicción, ésta debe hacerse de tal manera que se tenga presente la acción del azar y, así entonces, la predicción debe ser con probabilidad. En el ejemplo que di anteriormente, el problema consiste en saber qué tiene que hacer nuestro profesor para efectuar esas predicciones.

Stadi Shka: Sí; ese es otro problema. Me imagino que deberá tener una idea de cómo se ha comportado el sistema en ocasiones pasadas.

Chepa: Pero eso significa que nuestro pobre profesor tendría que mirar al mismo tiempo todas las notas de todos los parciales de todos los estudiantes del curso. Eso me parece muy difícil. Más aún si tenemos en cuenta que de las tres neuronas que él tiene, una está dañada, la otra está normalmente dormida y la otra se fue de vacaciones...

Ana Liza: Entonces, el profesor tendría que hacer algún tipo de resumen de todos esos datos para poder comprenderlos. Y su resumen debería basarse sólo en una parte de todo lo que podría observar. Si no lo hiciera así, si lo quisiera hacer con base en todas las observaciones, entonces no tendría sentido predecir; en ese caso sería una descripción lo que estaría realizando. De todas maneras, trabajar con todas las posibles observaciones, en muchos casos, resulta físicamente imposible y casi siempre muy costoso.

Stadi Shka: En realidad, no sé cómo va él a resumir esos datos. Pero supongamos que él sí sabe. ¿Qué hará después de resumir la información?

Ana Liza: Una vez que tenga resumidos los datos, tendrá que utilizarlos adecuadamente e interpretarlos para extraer de ellos generalizaciones y conclusiones que son las que permitirán que el profesor haga predicciones sobre lo que espera que suceda en ocasiones similares a las que ha analizado; y claro,

esas predicciones no estarán libres de error, pero lo importante es que ese error sí se puede cuantificar, se puede medir de alguna manera.

Stadi Shka: Y, frecuentemente el proceso no termina en la predicción, sino que se extiende hasta el punto de tomar decisiones. Es más, se hacen predicciones por la necesidad que existe de tomar decisiones. Frente a un determinado problema social que tiene diversas alternativas de solución, cuando se ha hecho una predicción, se pueden evaluar tales alternativas de solución y a partir de esa evaluación se puede tomar una decisión.

Ana Liza: En el caso de las calificaciones del parcial de nuestro curso, la decisión crucial que debe tomar el profesor es hacernos un examen práctico. ¿Verdad, Chepa?

Chepa: Todo muy bonito, pero muy abstracto. Porque, ¿quién le va a decir, o nos va a decir, cómo hacer esos resúmenes y esas predicciones?

Ana Liza: Pues ahí está el meollo del asunto. En lo que queda del semestre nos vamos a dedicar a aprender técnicas para resumir y para predecir en sistemas sociales. Y el conjunto de esas técnicas es lo que se llama *estadística*.

Chepa: Como quien dice, vine aquí para que me explicaran lo que no entiendo y ustedes, en cambio, me cuentan que hay muchas cosas que no sé y que voy a ver en el resto del semestre. ¡Muchas gracias! ¡Dizque amigas!

a. ¿Qué significa simplificar un problema social? ¿Qué relación hay entre un problema social y el correspondiente sistema social?

b. ¿Qué tan seguras son las generalizaciones obtenidas de la parte al todo en los sistemas sociales?

c. ¿Por qué debe tenerse presente la acción del azar en los sistemas sociales? ¿Cuál es la relación entre el azar y la probabilidad?

d. A continuación se dan tres situaciones problemáticas. Escoja una de ellas y construya un sistema social que le facilite abordar y manejar el problema. (Haga explícitas todas las suposiciones que crea necesarias para poder construir el sistema social que se le pide.)

- Imagine que usted es el jefe de la sección de control de calidad de una cierta de fábrica y durante los últimos tres meses ha recibido, de distintos clientes, quejas con respecto a la calidad de los artículos producidos. Usted está interesado en analizar el problema para decidir qué políticas debe adoptar.
- Usted es el director de un colegio y tiene indicios de que los estudiantes del colegio están consumiendo drogas alucinógenas. Usted está interesado en analizar el problema para decidir qué políticas debe adoptar.
- Usted es el jefe de la división de Bienestar Universitario en una cierta universidad y ha notado que la asistencia de los estudiantes a las actividades culturales y de recreación que programa la división es muy reducida. Usted está interesado en analizar el problema para decidir qué políticas debe adoptar.

Ejemplos de sistemas sociales

Introducción

Este capítulo tiene básicamente dos propósitos: uno, aclarar algunas afirmaciones hechas en los dos capítulos anteriores; y otro, presentar de manera informal una lista de puntos que deben tenerse en cuenta siempre que se vaya a realizar cualquier investigación estadística.

Se presentan, como ejemplos, tres problemas sociales, de diferente contenido, con una definición de sus respectivos sistemas sociales. A través del primer caso, además de presentar en qué consiste la construcción del modelo, se determina una lista de pasos que constituyen una correcta aproximación al problema que se tiene que resolver y que necesariamente deben haberse dado antes de pretender aplicar la estadística a la solución del problema. Además, se presentan dos situaciones concretas con sus correspondientes preguntas las cuales dan la oportunidad de completar o mejorar el modelo construido. Finalmente, se propone un ejercicio en cuya solución se puede observar la comprensión de los conceptos mencionados en este capítulo y los dos anteriores.

Primer ejemplo: Dime cómo vistes y te diré qué estudias

(Diálogo en la plazoleta exterior de la cafetería.)

Ana Liza: Hola Stadi Shka, ¿qué haces ahí sentada? Se diría que llevas un buen tiempo sin hacer nada.

Stadi Shka: Por el contrario, Ana Liza. Estoy trabajando para mi curso de estadística.

Ana Liza: ¿Trabajando? A mí me parece que estás disfrutando del "paisaje". Mira: allá viene Juan Mario, el estudiante de ingeniería que nos gusta tanto.

Stadi Shka: ¿Cómo sabes que estudia ingeniería? Solamente lo hemos visto de lejos un par de veces. O, ¿será que tú ya has estado haciendo tus averiguaciones?

Ana Liza: Realmente no estoy segura de que estudie ingeniería. Lo que pasa es que basta verlo para intuirlo. Mírale los pantalones bien limpiecitos, el suéter de rombos y la calculadora colgada de la cintura. No me vas a decir que estudia filosofía...

Stadi Shka: Tienes razón, Ana Liza. Es muy probable que Juan Mario estudie ingeniería. Y, precisamente eso tiene que ver con el trabajo que estoy haciendo desde hace dos horas, aquí sentada.

Ana Liza: No me vengas con el cuento de que estabas trabajando. Insisto en que estabas sencillamente mirando el paisaje.

Stadi Shka: Pero es que mirar el paisaje hace parte de mi trabajo para hoy. Figúrate que nuestro profesor de estadística nos ofreció la oportunidad de ganarnos unos puntos adicionales...

Ana Liza: ¡Unos puntos adicionales! Tu profesor sí que es buena persona. El nuestro nos tiene estudiando para un examen que haremos mañana. Y, ¿de qué se trata el trabajo?

Stadi Shka: Nuestro profesor propuso darnos unos puntos adicionales, si somos capaces de adivinar qué carrera estudia un estudiante de la Universidad, basándonos en su forma de vestir. ¡Ah! y podemos hacerle al estudiante, máximo una pregunta si ésta no tiene relación con su carrera. Claro está, que cualquiera que sea nuestra respuesta debemos justificarla, estadísticamente. Por eso, llevo dos horas mirando pasar a la gente.

Ana Liza: Y, ¿tú crees que mirar a la gente te va ayudar a resolver el problema?

Stadi Shka: Me da la impresión de que no mucho. Sin embargo, esta pequeña observación me ha ayudado a acercarme un poco al problema y creo que ya tengo una idea de cómo comenzar mi investigación.

Ana Liza: Y, ¿qué es lo que crees que tienes que hacer?

Stadi Shka: Lo primero, *definir el problema.* Y, como éste es claramente bastante complejo, voy a tratar de *identificar el contexto* dentro del cual se enmarca, es decir, voy a delimitar de la manera más completa y clara posible a qué seres se referirá mi investigación.

Ana Liza: Mi profesor de estadística nos ha insistido mucho en la importancia de aproximarnos a los problemas complejos a través del concepto de sistema social. ¿Cuál crees que es el sistema social?

Stadi Shka: Por un lado, pienso que el problema se restringe a los estudiantes de la Universidad, es decir, el contexto en el que se enmarca el problema es el conjunto de todos los estudiantes que están inscritos en la Universidad este semestre. Y por otro lado, para construir un sistema social que corresponda al problema tengo que identificar los elementos relevantes del problema y descubrir las interrelaciones que existen entre esos elementos. Sin embargo, imagínate que estuve averiguando cuántos estudiantes hay en la Universidad y ¡son más de 6.000!

Ana Liza: Eso quiere decir que no puedes pretender observar a todos y cada uno de los estudiantes.

Stadi Shka: Por supuesto. Y además, antes de lanzarme a tratar de comprobar si verdaderamente existe una relación entre la forma de vestir del estudiante y la carrera que éste estudia, quisiera tener una idea general del comportamiento de estos elementos. Para darte un ejemplo, quisiera saber si es razonable pensar que el semestre que cursa un estudiante es un factor externo que influye en el problema. Pero no puedo preguntarles a todos para sacar la distribución real de frecuencias de toda la *población*. ¿Qué hago?

Ana Liza: Pienso que debes comenzar por organizarte un poco, intentando hacer una primera aproximación a la construcción de un sistema social para el problema. Y para lograr eso basta, por ahora, identificar los factores externos y los elementos que son más relevantes para el problema en cuestión. ¿Se te ocurren algunos?

Stadi Shka: Claro. Ya te había comentado que me interesaría conocer el semestre en el que está cada estudiante. Además, es claro que tengo que conocer la forma de vestir y la carrera que estudia, puesto que allí está el centro del problema. También creo que me interesaría conocer alguna medida de los ingresos de la familia. En resumen tengo los siguientes elementos: (*anotando*)

forma de vestir, carrera, semestre, ingreso familiar.

Ana Liza: Pues ya tienes una base sobre la cual construir un sistema social. Sin embargo, recuerda que un sistema social no es únicamente una lista de elementos. Es esta lista de elementos, junto con unas interrelaciones entre ellos. ¿Se te ocurren algunas posibles interrelaciones?

Stadi Shka: Por supuesto. Pienso que *semestre* y *forma de vestir* están relacionados así:

> entre más adelantado esté el estudiante en su carrera, más seriamente se vestirá o usará la ropa que impone la "moda profesional".

Por otra parte, no estoy segura de que el nivel de ingresos de la familia pueda llegar a ser importante en la forma de vestir del estudiante. Ana Liza, ahora tengo otro problema.

Ana Liza: ¿Y, ahora qué te pasa?

Stadi Shka: Ahora que tengo una idea más clara de lo que es el problema, dado que he construido un modelo que lo ha simplificado, se me han comenzado a ocurrir ideas de lo que puede suceder al interior de ese sistema social; pero, esas son puras sensaciones personales, ¡impresiones subjetivas! Y mi profesor de estadística no me va a aceptar un trabajo basado en "impresiones".

Ana Liza: Pero, es que ésa es precisamente la función principal de la estadística: darte bases con las que puedas justificar racionalmente tus hipótesis acerca de un sistema social y darte herramientas para que puedas resolver un problema que te hayan asignado o que te interese. Lo importante es que te des cuenta de que la estadística es una herramienta, y que esta herramienta puede cumplir apropiadamente con sus propósitos solamente en aquellos casos en los que el investigador se haya aproximado correctamente al problema que tiene que resolver. Como normalmente los problemas a los que el investigador se enfrenta son problemas complejos, esto significa que es necesario llevar a cabo un proceso previo antes de la aplicación de la estadística. Este proceso —que podemos llamar la construcción del "modelo del sistema social"— implica que el investigador:

- Identifique la *realidad* que va a investigar.
- Determine y exponga explícitamente los objetivos de su investigación.

- Delimite el sistema social al identificar qué factores se consideran externos al problema.
- Delimite el sistema social al identificar qué elementos se consideran relevantes de acuerdo a los objetivos de la investigación.
- Delimite el sistema social al determinar las interrelaciones relevantes entre los elementos identificados, de acuerdo a los objetivos de la investigación.
- Formule hipótesis acerca de estas interrelaciones y del comportamiento de los elementos en cuestión.

Solamente en este momento, cuando el investigador haya satisfecho las etapas anteriores, podrá intervenir la estadística como medio para manejar las dificultades que resultan de la presencia del azar y de la complejidad del sistema social que se está estudiando. Es decir, es en este momento cuando se requieren de herramientas estadísticas para resumir la información e interpretarla.

Stadi Shka: Me descresta tu sabiduría, Ana Liza. Y, entonces ahora, ¿qué hago con los 6.000 estudiantes de la Universidad? ¡Yo no puedo ir a averiguarle la vida a todos y cada uno de ellos!

Ana Liza: Aunque pienso que, dada tu personalidad, esa sería una actividad que no te desagradaría, tienes razón: no puedes pasarte un mes recogiendo toda esa información. Yo creo que tendrás que preguntarle a unos cuantos y ver si lo que te dice esa *muestra* realmente es válido para la población, o sea para todos los estudiantes.

Stadi Shka: Y, ¿qué condiciones debe cumplir esa muestra? ¿Cómo debo construirla?

Ana Liza: En realidad, no sé responderte a esas preguntas. Pero, estoy segura de que deben existir unas ciertas condiciones para conseguir una muestra sobre la cual se puedan basar válidamente las conclusiones acerca de la población. Posiblemente, esas preguntas que tú formulas ahora se resolverán en el curso más adelante.

Stadi Shka: Bueno... tengo clase, Ana Liza. Nos vemos.

Ana Liza: Adiós.

Un resumen

En el diálogo anterior nos encontramos con que Stadi Shka tenía que resolver, desde el punto de vista estadístico, una serie de preguntas acerca de una población. La dificultad que se le presenta a Stadi Shka en un primer momento consiste en identificar y definir el sistema social dentro del cual se enmarca el problema. Para poder conocer un poco el interior del sistema social, se desea saber cómo es el comportamiento general de algunos de sus elementos relevantes. Lo ideal sería considerar **toda** la población y obtener **toda** la información. Sin embargo, como sucede en la mayoría de las investigaciones estadísticas, esto es imposible desde el punto de vista práctico. Esto tiene que ver con el hecho de que toda investigación estadística cuenta con una cierta cantidad de recursos (físicos, financieros, computacionales) y esos recursos no son suficientes para analizar toda la población.

Además, se han introducido ya los primeros factores o consideraciones que intervienen cuando se comienza toda investigación estadística:

- La necesidad de *definir el problema* a través de la identificación del sistema social dentro del cual éste se enmarca.
- La determinación de los *objetivos* de la investigación que se deducen de la definición del problema y de la identificación del sistema social.
- La necesidad de *recoger información* para conocer el sistema social y hacer un análisis previo del mismo.
- La necesidad de *recoger la información* a través de una *muestra*, dado que, en general, los recursos disponibles no permiten recoger la información concerniente a toda la *población*.
- La necesidad de que la muestra cumpla algunos requisitos para efectos de la calidad de las conclusiones que se obtengan a partir de ella.

Segundo ejemplo[1]

Un grupo interdisciplinario formado por un psiquiatra, un terapista y un psicólogo deseaba estructurar un programa especializado, dirigido a la rehabili-

1 Este problema sobre la rehabilitación sexual del minusválido fue formulado por Marta Patricia Ulloque, estudiante de Psicología de la Universidad de los Andes.

tación sexual del minusválido. Para llevar a cabo su propósito establecieron contacto con Teletón con el fin de conseguir patrocinio y ayuda técnica y humana. El programa se inició hace cinco meses[2] y hasta ahora ha atendido un total de 250 personas, hombres y mujeres cuyas edades oscilaban entre 18 y 45 años, radicadas en la ciudad de Bogotá. Ellos se han incorporado más efectivamente al ejercicio de la vida sexual y social en general.

Problema de investigación. Evaluar la efectividad de un tratamiento dirigido a la rehabilitación sexual del minusválido, cuando éste es incorporado a su vida normal activa y a su rol social.

Objetivos. Algunos de los objetivos de la investigación se expresan de la siguiente manera:

- Plantear alternativas de solución para la rehabilitación sexual del minusválido.
- Determinar métodos de trabajo social interdisciplinario en el campo estudiado.
- Identificar las necesidades sexuales del minusválido para lograr una vida sexual activa.
- Encontrar estrategias concretas que contribuyan a fomentar la autoestima en el minusválido y a prepararlo para su vida sexual en pareja.

Elementos considerados en el problema. Se incluyen, entre otros, los siguientes elementos:

- Tipo de lesión y áreas corporales implicadas
- Causa de la lesión e historia clínica del minusválido
- Edad en que se presentó la lesión
- Sexo
- Clase de tratamiento previo
- Experiencia sexual anterior
- Nivel o estrato socio-económico del minusválido

Interrelaciones entre los elementos. Considerando los elementos anteriormente mencionados, se pueden generar las siguientes interrelaciones:

- El tipo de lesión, por un lado, está muy relacionado con la posible mejoría que tenga el minusválido. Una lesión total como una parálisis

2 La fecha en que se escribió este texto fue mayo de 1991.

desde el cuello, implica un proceso más lento, costoso y difícil de orientar, a diferencia de una lesión parcial como una hemiplejia la cual es más sencilla, rápida y eficaz de tratar.

- La edad en la cual ocurrió la lesión puede influir en la percepción que se tiene de la sexualidad. Un minusválido muy joven puede estar en desventaja en cuanto al tratamiento en el ciclo de desinhibición y confrontación con su situación sexual si no le es posible evocar ninguna relación sexual o sensaciones relacionadas.
- El nivel socio-económico puede influir notablemente puesto que algunos interesados en su rehabilitación no están en capacidad de costearlo.
- El tratamiento de rehabilitación sexual de un individuo se puede ver afectado por alguna experiencia frustrante en tratamientos precedentes.
- El sexo del minusválido puede estar relacionado con la forma de reaccionar ante el programa como tal, ya que por tradición las mujeres han sido más recatadas y conservadoras en su apertura a la vida sexual.
- La edad actual del minusválido y sus necesidades sexuales pueden determinar el interés del minusválido hacia su propia rehabilitación y por consiguiente el éxito del programa.
- El éxito del programa depende del tipo de relación del minusválido con su pareja. La tolerancia, la aceptación de la lesión y de su gravedad, y el apoyo por parte de la pareja son factores determinantes del éxito del tratamiento en cada caso particular.

Restricciones de la investigación. Al igual que en toda investigación existen condiciones que limitan los alcances de ella. Para el particular, entre otros se consideran las siguientes restricciones:

- Nivel educativo de la pareja con respecto a temas concernientes a la sexualidad.
- Posibles conflictos interpersonales de tolerancia, comprensión y ayuda como pareja.
- Falta de recursos para seguir el tratamiento.

Tercer ejemplo[3]

El problema del sicariato en la juventud antioqueña es un fenómeno preocupante no sólo por el proceso de descomposición social que lo causó, sino también por la violencia que ha generado. Los jóvenes pertenecientes a los sectores menos favorecidos de Medellín han desarrollado una manera fácil y rápida de satisfacer su mentalidad de lucro: el asesinato de personas a cambio de sumas de dinero que van desde los miles hasta los millones de pesos.[4]

Dentro de los muchos factores (políticos, sociológicos, antropológicos, económicos, etc.) que de alguna manera se relacionan con la "cultura de la muerte" se encuentran los lingüísticos, ya que uno de los componentes más destacables de la cultura sicarial es el lenguaje. Las formas lexicales que estos jóvenes han conformado para referirse a sus actividades reflejan las concepciones que ellos tienen sobre la realidad y el entorno en el cual se desenvuelven. El análisis del fenómeno anteriormente planteado, por medio de la sociolingüística[5], permite identificar la función que cumple un lenguaje determinado dentro de un contexto social específico.

Con miras a realizar una investigación sociolingüística sobre el sicariato, un grupo de académicos observó el lenguaje hablado por 30 jóvenes delincuentes de las comunas nororientales de Medellín, durante octubre de 1989. El objetivo general que perseguían los investigadores era establecer que la mentalidad social decadente de los jóvenes sicarios se identifica con un lenguaje propio que manifiesta los anti-valores de la "cultura de la muerte".

Las variables consideradas en la investigación y su importancia para el desarrollo de la misma son:

- Las actividades de los individuos, con las cuales se muestra el papel de éstos en el sistema social.
- El nivel de educación de los individuos que permite ver el concepto de sociedad que ellos se han formado.

3 Este y todos los problemas sobre socio-lingüística que se presentan en el texto fueron formulados por Paola Valero, estudiante de Lenguas y Ciencia Política de la Universidad de los Andes.

4 Salazar, Alonso. *No nacimos pa' semilla.* Bogotá: Cinep, 1990.

5 La sociolingüística es el estudio científico del lenguaje enmarcado dentro del entorno social en el que aquél tiene lugar.

- La identificación del lenguaje propio de los sicarios, la cual ayuda a establecer el código de comunicación usado entre ellos.
- El significado del léxico empleado que contribuye a descubrir la relación que estos individuos establecen entre la realidad social y su propia concepción de ella.
- Las diferencias entre el lenguaje de los sicarios y el de otras personas del resto de la sociedad, que facilitan el establecimiento de relaciones cultura-lenguaje.

Las relaciones que se encontraron entre las variables pueden resumirse así: la pertenencia de un individuo a una clase social determinada y su falta de educación hacen que las posibilidades de ascenso en la escala socioeconómica sean cerradas para esta persona. Por esto, el individuo, desde muy joven debe suplir sus expectativas mediante una ocupación que reporte fácilmente beneficios económicos. Además, el desarrollo de una actividad delictiva induce al individuo a manejar un lenguaje que, si bien refleja una serie de realidades muy concretas, no las nombra con las palabras estándar que les corresponden sino con un conjunto de convenciones que establecen una nueva relación entre significante y significado. Como resultado de lo anterior, se obtiene un lenguaje con características semánticas, lexicales y morfológicas muy particulares. Tal lenguaje refleja las situaciones sociales de una cultura diferente a la tradicional antioqueña capitalina.

Construya usted el sistema social (1)[6]

El clientelismo es un fenómeno característico de las democracias representativas. Sin lugar a dudas existen diferentes concepciones acerca de su significado: "Para unos se refiere al nombramiento de funcionarios públicos incompetentes; para otros, a la compra de votos y al tráfico de influencias; o a la inmoralidad imperante en la política y en la administración pública".[7]

Para estudiar como investigadores sociales un fenómeno como el del clientelismo es preciso hacer algunas consideraciones históricas de importancia.

6 Este y todos los problemas sobre Ciencia Política que se presentan en el texto fueron formulados por Rocío Mariño, estudiante de Ciencia Política de la Universidad de los Andes.

7 Eduardo Díaz. *El clientelismo en Colombia.*

Según algunos autores, el período de cimentación de los partidos liberal y conservador, en Colombia, se vio enmarcado por el sentimiento de pertenencia a cada una de las dos facciones, lo que contribuyó significativamente a que tales partidos se configuraran y consolidaran como entes capaces de canalizar las demandas de la comunidad.

Con la instauración del Frente Nacional "las gratificaciones emocionales partidistas fueron truncándose por la búsqueda de algún favor burocrático o económico como condición de fidelidad partidista".[8] Debido a la escasez de bienes y servicios del Estado institucional, *el compadrazgo* y el *favor personal* se constituyeron en mecanismos de supervivencia comunitaria a nivel regional, lo que conllevó a que el clientelismo se convirtiera en "articulador político" de grupos y clases sociales. De esta manera vamos viendo cómo el sentimiento de pertenencia partidista es sustituido por la necesidad de afiliación a alguno de los partidos como condición para aspirar a beneficios económicos. Es preciso señalar que esta despolitización de los partidos se generó especialmente por el control de los movimientos políticos a nivel regional.

Como investigadores sociales nos cuestionamos acerca de la presente coyuntura política nacional y deseamos analizar el desarrollo de las recientes alternativas para salir de la crisis del régimen político bipartidista.

Indudablemente los últimos gobiernos han presentado propuestas que de alguna manera buscan reducir el clientelismo en los municipios colombianos como mecanismo deslegitimador del sistema. Por tal razón, conviene resaltar la importancia de la Elección Popular de Alcaldes (EPA) como reforma descentralizadora dentro del proceso electoral colombiano. Es necesario tener en cuenta ciertos criterios que nos permitan ver, como investigadores sociales, si la Elección Popular de Alcaldes en los municipios colombianos[9] logró romper con la forma tradicional de hacer política (clientelismo) y que, además, nos permitan establecer si el proceso (EPA) se convirtió en un verdadero mecanismo de representación y puente de comunicación entre los ciudadanos y el Estado.

8 Francisco Leal Buitrago. *Estado y política en Colombia.*

9 Vale la pena señalar que debido al alto número de municipios colombianos, debemos basarnos en el análisis de los municipios aledaños a Bogotá, en el período comprendido entre marzo de 1988 y marzo de 1990. Por tanto, las conclusiones a las que lleguemos sólo podrán ser aplicables a aquellos municipios colombianos que hacen parte de la población de estudio.

A continuación se consideran algunos elementos que son relevantes para lograr una primera aproximación al fenómeno del clientelismo en relación con la EPA:

- **El potencial electoral**. Está relacionado en gran medida con el nivel de abstención. Con base en él se puede observar el comportamiento de la participación electoral en la muestra escogida.
- **La situación económica del municipio**. Es relevante porque se relaciona con el nivel de vida de los habitantes del municipio, factor que permite ver en qué grado el clientelismo ha influido en la vida municipal, pues con base en el grado de desarrollo se puede observar qué tan alto es el manejo económico de los gamonales de la región.
- **La orientación partidista de los ciudadanos**. Es importante tenerla en cuenta ya que en la mayoría de los casos, ésta se ve afectada por factores clientelistas que incitan al ciudadano a votar a cambio de algo y no por su identificación con dicha organización.
- **La orientación partidista de los candidatos**. En la medida en que el candidato sea capaz de transmitir la ideología concreta de su partido, de alguna manera logrará captar más votos a favor de su organización.

Vemos de esta manera, la necesidad que tenemos como investigadores sociales, al afrontar el estudio de un fenómeno determinado, de considerar los siguientes factores: definir el problema, ubicándolo dentro de un contexto de tiempo y lugar determinados, resaltando la importancia del fenómeno para una determinada situación (en el caso anterior el clientelismo y el sistema electoral colombiano). A partir de lo anterior, distinguir los elementos que a nuestro modo de ver son básicos y presentan entre sí relaciones relevantes para el estudio que interesa y que a la postre servirán de postulados para formular las hipótesis, comprobarlas mediante la investigación y llegar a una conclusión.

a. Identifique el problema de estudio. (Recuerde que como parte de la definición del problema está la delimitación del correspondiente contexto).

b. Señale los objetivos de la investigación y mencione otros que para usted sean importantes.

c. ¿Hay elementos relevantes, según su criterio, que no hayan sido considerados al definir el sistema social correspondiente al problema que se

está tratando? Exprese cuáles son esos elementos y diga por qué son relevantes.

d. Según su criterio, ¿hay entre los elementos, interrelaciones relevantes que no hayan sido consideradas al definir el sistema social correspondiente al problema que se está tratando? Exprese cuáles son. Justifique su respuesta.

e. Haga una propuesta sobre cómo conformaría una muestra a partir de la cual pudiera recoger información sobre toda la población. Explique claramente.

f. Haga una propuesta sobre cómo obtendría la muestra. Explique.

Construya usted el sistema social (2)[10]

Según dicen por ahí, los mariscos tienen un efecto afrodisíaco en los seres humanos a determinada edad. También se sabe que todo elemento afrodisíaco es un excitante sexual que al ser ingerido, aplicado o utilizado produce en el organismo ciertos cambios tales como aumento en la frecuencia cardíaca y respiratoria, dilatación de la pupila y "piel de gallina".

Para comprobar esta hipótesis un estudiante de psicología se fue un sábado por la noche al restaurante La Fragata (ya que los sábados es cuando hay una mayor diferencia de edades: aproximadamente entre 18 y 60 años) y preguntó en cada una de las mesas si aceptaban hacer un pequeño experimento a la salida. Sólo algunos aceptaron.

a. Defina el problema. (Recuerde que como parte de la definición del problema está la delimitación del correspondiente contexto.)

10 Este y todos los problemas sobre Psicología que se presentan en el texto fueron formulados por Paola Turbay, estudiante de Psicología de la Universidad de los Andes.

b. Señale los objetivos de la investigación y mencione otros que para usted sean importantes.

c. ¿Cuál es la muestra que se ha escogido para estimar ciertos datos acerca de la población? ¿Por qué se escogió esa muestra y no otra?

d. ¿Cuáles son los elementos más relevantes del problema y qué relación existe entre ellos?

e. ¿Qué problemas, cree usted, que se podrían presentar en el experimento?

Invente su propio problema

1.- Proponga un problema de la "vida real" (preferiblemente relacionado con cuestiones de la Universidad) y lleve a cabo los pasos siguientes:

a. Defina el problema.

b. Plantee los objetivos de su investigación.

c. Determine el sistema social correspondiente. Esto es, diga cuáles son los elementos relevantes y las interrelaciones relevantes entre esos elementos.

d. Determine cuáles son las restricciones de recolección de información.

e. Haga una propuesta sobre cómo conformaría una muestra a partir de la cual pudiera recoger información sobre toda la población. Explique claramente.

f. Haga una propuesta sobre cómo obtendría la muestra. Explique.

Población y muestra

Introducción

En el capítulo anterior se mencionaron dos conceptos básicos en la estadística: población y muestra. En este capítulo se definen de manera precisa y se hacen algunos comentarios que se deben tener en cuenta cuando haya que manejar dichos conceptos. Aunque nuestra intención no es entrar en detalles sobre la teoría del muestreo, se presenta una lectura con la cual se logra tener una idea de los problemas que pueden surgir en una investigación cuando las muestras han sido tomadas de manera inadecuada. Finalmente, hay una sección de ejercicios.

Vamos al grano

El fin último de la estadística —por lo menos de la estadística inferencial— es hacer inferencias. Es decir, obtener generalizaciones, hacer predicciones, hacer estimaciones o determinar si una hipótesis dada se puede rechazar o no con algún grado de certidumbre. La estadística inferencial debe dar las herramientas necesarias para llevar a cabo esas tareas a partir de la información que arroje un conjunto "pequeño" de datos, y también debe dar las herramientas que hagan posible aplicar los resultados obtenidos a un cierto conjunto "más grande" de datos, de donde se supone que se extrajo el conjunto pequeño. El primer conjunto mencionado en el párrafo es una muestra y el segundo es una población. Puesto que los conceptos de población y muestra están estrechamente ligados con el fin último de la estadística, la perfecta comprensión de los correspondientes significados es una condición necesaria para manejar apropiadamente temas que se estudiarán posteriormente.

> La *población de una investigación* o simplemente la *población de estudio* se define como el conjunto de **todos** los entes a los cuales se pueden aplicar las conclusiones obtenidas a través de la predicción, estimación, o verificación de una hipótesis, acciones éstas realizadas como parte final de la investigación.

A partir de lo dicho anteriormente se deduce que la población de una investigación puede ser un conjunto de personas, un conjunto de animales o un conjunto de objetos. De qué tipo de ente se trate, no interesa; eso depende, por supuesto, del asunto alrededor del cual se esté haciendo la investigación. Lo que sí es muy importante es la precisión con que se defina la población para cada caso. Las características de esa definición son las mismas que tiene la definición de cualquier conjunto. En otras palabras, lo que se diga acerca de los elementos de la población debe ser aquella información que permita, en todo caso, decidir si un determinado ente es o no un elemento de la población. En términos generales, las características que se expresen para delimitar los elementos de una población establecen el contexto en el cual se va a trabajar, porque ubican a dichos elementos en el espacio y en el tiempo y además señalan las condiciones que están presentes y que es necesario tener en cuenta para cualquier análisis que se haga.

Veamos un ejemplo: un psicólogo quiere determinar si existe relación entre el rendimiento escolar y el hecho de que los niños provengan de familias de padres separados. Para ello toma un grupo de niños y les hace un seguimiento. ¿Cuál es la población de estudio? Tal como está expresado el problema tendríamos que aceptar que la población es el conjunto de todos los niños. Sin embargo, un estudio, en ese aspecto, que tenga como población a todos los niños del mundo no tiene sentido, por muchas razones. Por ejemplo, hay diferencias significativas entre la concepción de la vida familiar que tienen los latinos y la que tienen los norteamericanos; no se puede comparar el nivel de escolaridad de los países desarrollados con el de los países menos desarrollados; los resultados que se obtengan con una muestra tomada a finales del siglo XX no se pueden comparar con los que se habrían obtenido a comienzos del mismo siglo, etc. Por tanto, si se quiere hacer una investigación seria que arroje alguna información válida es necesario limitar el alcance de los resultados. Por ejemplo, el problema podría replantearse así: un psicólogo quiere determinar si existe relación entre el rendimiento escolar y el hecho de que los niños provengan de familias de padres separados. La investigación se va a realizar para los niños de Colombia que están entre 4 y 7 años. La investi-

gación se hace con la intención de determinar si es conveniente darles un tratamiento especial antes de que inicien la primaria. Además esta investigación tendrá un alcance de 10 años porque se espera que las situaciones esenciales involucradas en este asunto se mantengan iguales por lo menos dentro de los próximos 10 años. Aunque así replanteado el problema, se ha limitado en gran medida el contexto, de todas maneras la población aún sigue siendo muy vasta, por ejemplo, para el caso en que ese fuera el tema de tesis de un estudiante universitario, pues el costo de tal investigación sería muy alto.

Ahora considere que lo que se quiere estudiar es la relación que hay entre el tipo de alimentación que reciben los niños y la cantidad de horas que ellos duermen diariamente. También suponga que la investigación se va a realizar para los niños de Colombia que están entre 4 y 7 años. En este caso y en el anterior se trata de la misma población de estudio. Sin embargo, es evidente que el tipo de mediciones que se harán sobre los elementos que conformen las muestras en cada caso es diferente.

Con el ejemplo anterior se quiere poner de manifiesto el siguiente hecho: a un mismo conjunto de elementos puede haber asociados muchos conjuntos de datos, puesto que en los elementos de una población de estudio, se pueden observar características muy variadas. Conviene por tanto hacer alguna precisión al respecto.

> La *población de datos* o simplemente la *población* se define como el conjunto de **todas** las mediciones que es posible obtener a partir de observar una cierta característica en cada uno de los elementos de la población de estudio.

De lo dicho anteriormente se deduce que para un cierto estudio habrá tantas poblaciones de datos como elementos relevantes se quieran considerar. Además, los elementos de dichas poblaciones serán, o bien valores numéricos, o bien valores cualitativos.

Veamos un ejemplo: para el caso de la relación entre el tipo de alimentación que reciben los niños y las horas que duermen diariamente, habría dos poblaciones de datos. Una, corresponde a las mediciones obtenidas a partir de observar en todos los niños de Colombia, que están entre 4 y 7 años el tipo de alimentación que reciben. Suponiendo que ésta se califica como buena, regular o deficiente, la correspondiente población de datos sería el conjunto constituido por la calificación, en ese aspecto, de cada uno de los niños. La otra población de datos corresponde a las mediciones obtenidas a partir de obser-

var en todos los niños de Colombia, que están entre 4 y 7 años el número de horas que duermen diariamente. En este caso, la población de datos sería el conjunto constituido por el número de horas de sueño de cada uno de los niños.

Una vez que se han comprendido los conceptos población de estudio y población de datos, es fácil entender los correspondientes conceptos de muestra de estudio y muestra de datos.

> Una *muestra de estudio* es cualquier subconjunto no vacío de la población de estudio. Y, una *muestra de datos* es cualquier subconjunto no vacío de la población de datos.

En resumen: vamos a distinguir el conjunto de entes sobre los cuales efectuamos observaciones, del conjunto de mediciones obtenidas a partir de las observaciones que se hagan sobre tales entes. Y son los conjuntos de mediciones con los que trabajaremos para hacer análisis.

a. Para cada una de las siguientes situaciones, usted debe describir la población de estudio de la cual usted seleccionaría una muestra.

- Se quiere hacer una estimación del consumo mensual de agua en su casa.
- Se quiere hacer una estimación del tiempo que emplea Fernando en llegar desde su casa a la Universidad, cuando viaja en bus.

Algo más acerca de las muestras[11]

A continuación se plantean cinco situaciones descritas de manera muy breve, cada una con una corta alusión a aspectos que tienen que ver con la acción de obtener muestras de una población. Con ello se quiere destacar, de manera informal, la existencia de algunos problemas inherentes a la toma de muestras

11 Versión libre de "La muestra que presenta un factor de influencia en sí misma" en *Cómo mentir con estadísticas* de Darrell Huff.

y a su manejo; se quiere establecer requisitos que deben cumplir las muestras para que los informes que se derivan del estudio de ellas sean, en lo posible, válidos; y además, se quiere justificar la necesidad de una posición crítica y activa frente a las afirmaciones hechas a partir de estudios realizados con base en muestras de una población.

Ejemplo 1. En una encuesta que se hizo en las distintas zonas de la ciudad, con la finalidad de conocer el número de lectores de revistas y determinar preferencias, la pregunta clave era: "¿Qué revistas leen los miembros de su familia?"

¿La manera de obtener la información es la adecuada? Muy probablemente no. Tal vez esa encuesta revela solamente el grado de esnobismo de los informantes. Posiblemente si se desea saber lo que cierta clase de público lee, no se obtiene una información confiable a partir de una pregunta directa. Podría ser más efectivo, por ejemplo, visitarlos y decirles que se desea comprar todas las revistas viejas que tengan. Aun así, no se podría asegurar que las revistas que vendan sean precisamente lo que leen.

Ejemplo 2. En un informe se dice algo así como "el americano medio se cepilla los dientes 1,02 veces por día".

¿Qué refleja la afirmación? Dado que los anuncios publicitarios establecen que la aceptación social depende de cosas tales como una fragancia agradable en la boca, ¿puede creerse que el "americano medio" que no se lave los dientes responderá sinceramente a la pregunta del encuestador? Estos resultados estadísticos pueden tener significado para quien quiera solamente la opinión de la gente acerca de la higiene dental, pero no se puede deducir mucho más.

Ejemplo 3. Un psiquiatra informó una vez que prácticamente todo el mundo está neurótico.

¿A quiénes se refiere la afirmación? Aparte del hecho de que esta afirmación destruye el significado de la palabra "neurótico", examinemos la muestra utilizada por el doctor. Es decir, ¿a quién observó el psiquiatra? Evidentemente partió del estudio de sus pacientes, que distan mucho de constituir una muestra de la población. Si un hombre fuera normal, él no sería paciente de tal psiquiatra.

Ejemplo 4. Para una encuesta de opinión, una entrevistadora consideró la estación del ferrocarril como el lugar ideal, ya que "allí se encuentra a toda clase de personas".

¿Se obtienen muestras representativas de la población? No, porque por ejemplo, las madres con hijos pequeños tienen una representación muy exigua en ese lugar.

Ejemplo 5. El promedio de los componentes de la promoción de 1924 de la Universidad de Yale gana 25.111 dólares al año.

¿Se puede confiar plenamente en el dato? El conocimiento exacto de los ingresos de una persona es muy difícil de lograr, a menos que ellos provengan exclusivamente de su salario y que el investigador tenga acceso a la nómina de empleados. Además, pocas veces, los ingresos del orden de 25.000 dólares provienen totalmente de un salario; la gente que se encuentra en este nivel disfruta, probablemente, de inversiones bien distribuidas. Por tanto, es muy probable que este dato haya sido obtenido de lo que dijeron los graduados de Yale. En una encuesta no es muy probable que una persona declare con veracidad cuánto gana, bien sea porque no desea tener problemas de impuestos, o porque siente vergüenza de declarar su verdadero ingreso. Es posible que las tendencias de sobreestimar y de subestimar los ingresos se neutralicen en una muestra grande, pero esto tampoco parece muy probable, según la experiencia que uno tiene del comportamiento de las personas y de la manera como la riqueza está distribuida.

¿Cómo se obtuvo ese dato? El dato deriva de una muestra. Es poco creíble, por no decir completamente inaceptable, que la encuesta se haya realizado entre todos los egresados de la promoción mencionada, pues al cabo de 25 años (según la fecha del artículo), es prácticamente imposible localizar a todos los miembros vivientes de la promoción. Además, no todos los que hayan recibido la encuesta habrán respondido; entre estos últimos estarán seguramente aquellos que no han obtenido éxito económico.

¿Es representativa esa muestra? Como ya se dijo, no es absurdo pensar que de aquellos cuya dirección se conocía, sólo algunos respondieron la encuesta, por tanto es evidente que la muestra ha omitido grupos que muy probablemente reducirían la media. Suponiendo que 25.111 dólares es una cifra representativa, lo que representa es, sin duda, el grupo especial de miembros de la promoción de 1924 cuyas direcciones se conocen y que están dispuestos a colaborar y publicar los ingresos de que disfrutan. Incluso en tal caso, hay que partir del supuesto de que estos hombres dicen la verdad.

En los ejemplos trabajados anteriormente hay implícitos conceptos tales como: muestra aleatoria, muestra aleatoria estratificada, técnica de muestreo,

muestra representativa y factores de influencia de una muestra. Todos esos conceptos son de gran importancia en la estadística; nosotros vamos a aproximarnos de manera rápida e informal a ellos.

La base de la técnica del muestreo se presenta con el ejemplo que se da a continuación: si se tiene una bolsa de fichas rojas y blancas existe sólo una manera de averiguar exactamente cuántas hay de cada color: ¡contarlas! Sin embargo, puede averiguarse de una forma más fácil, pero con aproximación, la cantidad de fichas rojas: se saca de la bolsa un puñado de fichas, se cuentan las que lo componen y se da por sentado que la proporción será la misma en todo caso. Si la muestra es lo bastante extensa y bien seleccionada, representará al conjunto con bastante aproximación en la mayoría de los casos. Si no es así, puede ser menos precisa que una hipótesis sensata, y no tiene nada que la apoye, a no ser un aire aparente de precisión científica. Es una triste verdad el hecho de que detrás de lo que leemos o creemos saber están las conclusiones derivadas de muestras subjetivas y demasiado pequeñas, o ambas cosas a la vez.

La muestra básica es la llamada *muestra aleatoria*, la cual se selecciona por azar, partiendo de una "población". Se toma, por ejemplo, uno de cada diez nombres buscados en un fichero o relación. De un sombrero se extraen cincuenta papeletas dobladas. Se entrevista a una de cada veinte personas que pasan por una determinada calle de Bogotá (pero tenga presente que esta no sería una muestra de la población del mundo, ni de los Estados Unidos, ni de Bogotá, sino solamente de la gente que pasa por tal calle en aquel momento).

Para comprobar que una muestra fue tomada al azar se procede según el criterio siguiente: ¿Tiene cada nombre o cosa de la población la misma probabilidad de formar parte de la muestra?

La muestra escogida al azar es la única que puede examinarse con completa confianza por medio de la teoría estadística, pero existe un factor desfavorable a la misma. Es tan difícil y cara de obtener, en muchos casos, que queda descartada por su costo. Un sustituto más económico, usado en todo el mundo para trabajos tales como las encuestas de opinión y estudios de mercado, es la llamada *muestra al azar por estratos*.

Para obtener esta muestra estratificada se divide la población en estratos (grupos bien diferenciados de la población) y de cada uno de ellos se saca, aleatoriamente, un grupo cuyo número de elementos sea proporcional al tamaño del estrato del cual provino.

En una *muestra aleatoria representativa* se incluyen, proporcionalmente, elementos de todos los diferentes grupos que haya en la población. No todas las muestras aleatorias son representativas.

Con frecuencia se publican artículos en las revistas, en los cuales se presentan afirmaciones respaldadas con encuestas o con reportes sobre muestras supuestamente representativas de una población, para estimular la credibilidad del lector. Al leer esa clase de artículos debemos poner en guardia nuestro sentido crítico, si no queremos tragar entero todo lo que se nos presente. Conviene pues, examinar dos veces lo que se lee, y evitar creer una cantidad de cosas que no son verdad.

Son varios los elementos que entran en juego y determinan la seriedad, la validez y la confiabilidad de un informe estadístico obtenido a partir de muestras. Se pueden mencionar aspectos relacionados con:

- La conformación de la muestra (tamaño, proporción en la que se representa cada uno de los grupos de la población en la muestra, aleatoriedad, técnica de muestreo utilizada para obtenerla)
- La calidad de la muestra (representatividad)
- La forma de obtener la información requerida (encuestas, entrevistas, preguntas directas, observación, etc.)
- La calidad de las respuestas dadas por los investigados (veracidad, autenticidad)

A continuación se presenta una lista de factores tendenciosos que pueden estar presentes en los resultados obtenidos a partir de muestras:

- Tendencia a dar respuestas agradables.
- Tendencia a responder lo que se acepta socialmente como óptimo.
- Tendencia a distorsionar ciertas realidades o a no hablar de ellas.
- Tendencia a explorar actitudes que pueden basarse en sentimientos o en prejuicios sociales, y que por tanto, arrojan información no auténtica.
- Tendencia a dirigirse a personas que poseen más dinero, mejor educación, mejor información, mejor aspecto, mejor comportamiento convencional, y unas costumbres más fijas que el promedio de la población que representan.
- Tendencia a aceptar la información que tiene visos estadísticos sin cuestionarla, sin cargarla de significado.

En todo caso, cuando los datos han sido filtrados a través de distintas fases de manipulación estadística y reducidas a una medida expresada en decimales, el resultado empieza a presentar una aureola de convicción que sólo se vería empañada por una revisión cuidadosa de la muestra.

Para que un informe basado en una muestra tenga valor, debe utilizar una muestra representativa, donde se hayan eliminado todos los posibles factores de influencia. Vale la pena tener en cuenta también, que la representatividad de una muestra puede ser destruida con la mayor facilidad, tanto por influencia de factores visibles como por la de los invisibles. Es decir, incluso en caso de que no pueda demostrarse que existe un factor de influencia apreciable, es prudente conservar cierto grado de escepticismo sobre los resultados, siempre que haya una posibilidad de influencia en alguna parte. Siempre la hay.

En realidad, no hace falta que la encuesta esté falseada, es decir, que se tergiversen los resultados deliberadamente a fin de crear una falsa impresión. La tendencia de la muestra a presentar un factor de influencia en el sentido que acabamos de explicar puede falsear la encuesta automáticamente.

a. Explique por qué es importante tomar muestras al hacer una investigación.

b. Explique en qué se basa la técnica del muestreo.

c. Realice un experimento que muestre en qué consiste la técnica del muestreo. Haga comentarios pertinentes a los resultados que obtiene.

d. ¿Qué es una muestra aleatoria?

e. Suponga que la población de un cierto estudio es el conjunto de alumnos que están inscritos este semestre en la Universidad, en Ciencias Sociales. Y el objetivo del estudio es establecer la relación que hay entre el tiempo que ellos dedican a la investigación y sus hábitos culturales. Es evidente que no es posible trabajar con la población, por tanto se requiere tomar una muestra. Sugiera métodos para obtener muestras aleatorias de la población. Además, justifique claramente por qué esos métodos conducen a muestras aleatorias.

f. Con respecto al problema definido en el item anterior, dé ejemplo de un método que conduzca a una muestra no aleatoria (las muestras no aleatorias se llaman sesgadas). Explique su respuesta.

g. ¿Qué es una muestra aleatoria por estratos? ¿Qué problemas se presentan al intentar hacer muestreo estratificado?

h. Con respecto al problema definido en el item **e)**, explique qué criterios podrían tenerse en cuenta para estratificar la población. Además, diga cómo obtendría una muestra aleatoria estratificada de la población.

i. ¿Qué condiciones debe cumplir la muestra para representar adecuadamente al conjunto del cual se extrae?

j. Explique por qué una muestra aleatoria puede no ser representativa. Dé un ejemplo.

k. Una caja contiene 100 bolas, de las cuales 36 son azules, 25 son negras, 9 son verdes y 30 son rojas. A continuación se dan 4 muestras tomadas de la caja que contiene 100 bolas:

Muestra	**Composición de la muestra**			
1	5 azules	5 negras	5 verdes	5 rojas
2	8 azules	4 negras	3 verdes	5 rojas
3	7 azules	5 negras	2 verdes	6 rojas
4	6 azules	5 negras	4 verdes	5 rojas

Determine cuál de esas muestras es la que más refleja la conformación de la población. Explique su respuesta.

l. En las siguientes situaciones planteadas, usted debe determinar cuáles de las técnicas de muestreo conducen a obtener muestras aleatorias y cuáles a obtener muestras no aleatorias o *sesgadas*.

- **Población de estudio**: residentes de Bogotá.
  Objetivo del estudio: se quiere determinar el grado de popularidad de un cierto candidato a la presidencia de la República.
  Técnica de muestreo: durante una semana se detiene a toda persona que pase por la esquina de la carrera 13 con la calle 60 y se le pide que responda a dos preguntas.

- **Población de estudio:** residentes en Tunja.
  Objetivo del estudio: se quiere estimar la calificación que se da a la prestación del servicio de teléfono.
  Técnica de muestreo: de la guía telefónica de la ciudad se selecciona un número de teléfono de cada 100, se llama allí y a quien conteste se le hace una encuesta que consta de 3 preguntas.

- **Población de estudio:** matas de lechuga en una huerta.
  Objetivo del estudio: se desea estimar el tamaño de las matas de lechuga en esa huerta.
  Técnica de muestreo: Extraer, con los ojos vendados, una mata de cada metro cuadrado.

- **Población de estudio:** estudiantes de la universidad donde usted estudia.
  Objetivo del estudio: usted quiere establecer si hay relación entre las calificaciones de un estudiante de la Universidad y la carrera que estudia.
  Técnica de muestreo: el miércoles de la semana entrante se seleccionará uno de cada veinte estudiantes entre los que ingresen a la Universidad por la puerta principal y le pedirá que responda una encuesta.

- **Población de estudio:** habitantes de Bogotá.
  Objetivo del estudio: un psicólogo quiere averiguar el nivel de esquizofrenia en Bogotá.
  Técnica de muestreo: el psicólogo se fue al restaurante Pozetto y le hizo un test a cada una de las personas que estaban allí.

- **Población de estudio:** estudiantes de la universidad donde usted estudia.
  Objetivo del estudio: se quiere conocer una cifra aproximada del nivel de drogadicción en la universidad.
  Técnica de muestreo: un estudiante escogió al azar a cinco personas de cada semestre de cada carrera y los sometió a un examen escrito.

- **Población de estudio:** estudiantes del colegio X.
  Objetivo del estudio: se quiere saber si el mal rendimiento de los estudiantes del colegio se debe a la forma en que los profesores dictan las clases.

Técnica de muestreo: el psicólogo de un colegio asistió a todas las clases de los alumnos del curso décimo, durante una semana.

- **Población de estudio:** estudiantes de la universidad donde usted estudia.

  Objetivo del estudio: el jefe del Departamento de Matemáticas de la universidad donde usted estudia, quiere saber si a los estudiantes les gusta la forma en que se están dictando todas las clases de matemáticas.

  Técnica de muestreo: un monitor del Departamento de Matemáticas se ubicó en una de las puertas de entrada a la universidad y le preguntó a 1 de cada 10 personas que salían, qué piensa al respecto.

Ejercicios

1.- La Comunidad Económica Europea (CEE) ha generado, sin lugar a dudas, una serie de expectativas a nivel internacional y particularmente a nivel latinoamericano. Un grupo de politólogos busca estudiar las repercusiones políticas de la consolidación del bloque económico europeo en América Latina y con base en su análisis, predecir si el proceso político "latino", con sus características específicas, estaría en capacidad de llevar a cabo un movimiento integracionista a nivel latinoamericano, teniendo en cuenta factores como costos y beneficios de integración.

El grupo de investigadores considera que es necesario analizar el problema teniendo como referencia una experiencia particular europea (cualquier país que conforme el grupo de los Estados de la Comunidad Económica Europea) en la que se contemplen elementos tales como: nivel de inflación, PNB, régimen político existente, clase de economía que posee el país (mercado, o, socialista) entre otros. Por supuesto es necesario considerar cada uno de esos elementos dentro de un cierto contexto; para este caso puede pensarse en 1990 como el año base del estudio.

a. Defina el problema de estudio.

b. Señale posibles objetivos de investigación.

c. Mencione qué otros elementos se podrían tener en cuenta para la investigación.

d. Determine la muestra con base en la cual se haría la investigación.

2.- La reforma de descentralización administrativa iniciada en el gobierno de Belisario Betancur y puesta en marcha en el gobierno de Virgilio Barco ha venido acompañada de procesos que como la Elección Popular de Alcaldes, incrementan la participación ciudadana y fortalecen la vida regional y municipal. Esto a largo plazo significa la modernización del Estado y hace que se cumpla cabalmente con las demandas de la comunidad. Un grupo de estudiantes de ciencia política realizó un estudio acerca del fenómeno de la Elección Popular de Alcaldes teniendo como referencia los municipios aledaños a Bogotá y analizando el caso particular del municipio de Chía, en el período comprendido entre 1988 (primera EPA) y 1990 (segunda EPA). Para construir un sistema social del problema que les interesaba analizar tuvieron en cuenta elementos tales como: participación de terceros partidos en el proceso, comportamiento tradicional del municipio; el clientelismo frente a la elección local de mandatarios; cambios en el voto (cautivo, opinión y de lealtad partidista).

a. Defina el problema.

b. Señale los posibles objetivos de la investigación.

c. Plantee algunas hipótesis sobre los elementos que el grupo de investigadores identificó como relevantes.

d. Y, usted ¿qué elementos relevantes cree que sería necesario considerar para estudiar el problema?

e. ¿Cuál es la muestra con base en la cual se va a analizar el problema?

3.- Un grupo de politólogos desea analizar el fenómeno de la Asamblea Nacional Constituyente en el departamento de Cundinamarca y para ello toman como referencia tres ciudades: Bogotá, Girardot y Zipaquirá. Los investigadores buscan determinar si esta convocatoria nacional por parte de las fuerzas sociales, políticas y económicas constituye realmente la salida a la honda crisis nacional que vive el país. Es decir, se busca determinar si la Asamblea Nacional Constituyente será capaz de construir un nuevo Estado y una nueva organización pública que interprete y exprese la reali-

dad política y social. Algunos de los aspectos que consideraron como relevantes para la investigación son: la democracia participativa o representativa; canales de participación (gremios, juntas de acción comunal, organizaciones indígenas); número de votantes en favor y en contra de la Asamblea; nivel de abstención.

a. Defina el problema que se pretende estudiar.

b. Señale cuál es la muestra de estudio.

c. ¿De qué manera cree usted que se podrían medir los elementos mencionados en el texto anterior?

d. Sugiera otros elementos relevantes para el análisis del problema en cuestión.

4.- Se rumora que la mayor concentración de consumo de droga en una cierta universidad se encuentra en los estudiantes de X carrera (no se sabe cuál). La psicóloga de la universidad debe averiguarlo para ofrecerle a dichos estudiantes unas sesiones de terapia de prevención y control.

a. ¿Cuál es el problema? (Recuerde que como parte de la definición del problema está la delimitación del correspondiente contexto.)

b. Señale los objetivos de la investigación y mencione otros que para usted sean importantes.

c. ¿Cuáles son los elementos más relevantes del problema y qué relación existe entre ellos?

d. Haga una propuesta sobre cómo conformaría una muestra a partir de la cual pudiera recoger información sobre toda la población. Explique claramente.

e. Haga una propuesta sobre cómo obtendría la muestra. Explique.

5.- Un psicólogo del colegio X de Bogotá cree que los alumnos del colegio que son hijos de padres separados tienen nivel académico medio-bajo (calificación 3.0). El psicólogo desea averiguar qué tanto coincide su apreciación con la realidad para poder brindar a tales estudiantes una atención especial.

a. ¿Cuál es el problema? (Recuerde que como parte de la definición del problema está la delimitación del correspondiente contexto.)

b. Señale los objetivos de la investigación y mencione otros que para usted sean importantes.

c. ¿Cuáles son los elementos más relevantes del problema y qué relación existe entre ellos?

d. Haga una propuesta sobre cómo conformaría una muestra a partir de la cual pudiera recoger información sobre toda la población. Explique claramente.

e. Haga una propuesta sobre cómo obtendría la muestra. Explique.

6.- Un fenómeno muy frecuente en las clases bajas de Bogotá es la elisión del sonido /k/ en combinaciones de sonidos consonánticos como /kt/, y /kf/ principalmente. Palabras como a**cc**ión, a**ct**uación o I**cf**es que deberían pronunciarse /aksión/, /aktuasion/ o /ikfes/ son pronunciadas corrientemente /asion/, /atuasion/ o /ifes/.

Con el fin de determinar las causas del fenómeno y su identificación como variante estigmatizada,[12] un grupo de lingüistas realizó un estudio durante mayo-septiembre de 1989 en los sectores de estrato 1 y 2 de la capital de la República de Colombia. Para ello se recolectó una muestra aleatoria de 1.000 individuos a quienes se les hizo una entrevista de quince minutos aproximadamente, sobre un tema que requería el uso de palabras con los sonidos en cuestión.

a. ¿Cuál es el problema de estudio?

b. ¿Cómo se delimita del problema?

c. ¿Cuáles son los objetivos de la investigación?

d. ¿Cuál es la población del estudio?

e. ¿Cuál es la muestra del estudio?

12 Una variante estigmatizada es un rasgo que se asocia con los grupos de baja posición social y con estilos informales de lenguaje.

f. ¿Cuál es la muestra de datos?

7.- Una de las técnicas de investigación más frecuentes en estudios sociolingüísticos es el análisis del discurso, herramienta que a partir de una muestra del lenguaje (oral o escrito) permite abordar un texto desde varios niveles: nivel del texto, nivel gramatical, nivel semántico, etc.

Usando esta técnica quiere mirarse el concepto de *poder* que se han formado los niños colombianos de acuerdo con las diferentes condiciones sociales en que se han desenvuelto. Para tal efecto, se recolectaron cartas de todos los niños que en el año 1989 cursaban quinto de primaria tanto en áreas rurales como urbanas, en escuelas públicas o privadas y en cualquier calendario escolar. Las cartas respondían a la pregunta: ¿Qué haría yo si fuera el presidente de Colombia?

a. ¿Cuál es el problema de estudio?

b. ¿Cuáles son los elementos importantes que se deben tener en cuenta? Explique su respuesta.

c. ¿Cuál es la población de estudio?

d. ¿Cuál es la muestra de estudio? ¿Es representativa de la población? Justifique su respuesta.

e. ¿Cree usted que la pregunta a la cual deben responder los niños es adecuada para adquirir información sobre lo que se quiere estudiar? Justifique su respuesta.

Variables

Introducción

Al igual que los conceptos de población y de muestra, el concepto de *variable* es bien importante en la estadística. En la siguiente frase se resume muy brevemente la relación que liga esos tres conceptos y por qué afirmamos que son básicos en el estudio de la estadística: la **estadística** permite hacer inferencia acerca del comportamiento de una **variable** en una **población**, a partir del análisis del comportamiento de dicha variable en una **muestra**.

En realidad, en todos los capítulos anteriores nos hemos referido al concepto de variable, aunque no hayamos mencionado exactamente tal palabra. Al hablar de los elementos relevantes en un problema, estamos refiriéndonos a las variables de interés para el caso. En este capítulo vamos a definir el concepto de variable y a establecer una clasificación de las variables según el tipo de valores que ellas asuman.

El apartamento de Perla Madonna

(Ana Liza se encuentra con su hermana, Perla Madonna, en una cafetería.)

Perla Madonna: Pues sí, Ana. Voy a comprar un apartamento en Bogotá y quiero que tú me ayudes a tomar la decisión. Desde hace algunos días he venido mirando los avisos clasificados que salen en el periódico *El Desinformador* y me parecen muy completos, de manera que voy a seguir haciéndolo, hasta completar un mes, para tratar de encontrar el apartamento que busco.

Ana Liza: Claro que te ayudo. ¡Qué dicha que puedas comprar tu apartamento! ¿Tienes aquí el periódico de hoy?

Perla Madonna: Sí. Tómalo.

(*Encuentra la página de avisos clasificados y la ojea rápidamente; todos los avisos que ve señalados anuncian apartamentos nuevos.*)

Ana Liza: ¡Uf! ¡Qué cantidad de apartamentos nuevos que ofrecen!

Perla Madonna: Es verdad; hoy ofrecen 30 apartamentos todos ubicados en Bogotá y nuevos. Yo quiero un apartamento para estrenar, no me importa que sea pequeño, pero lo quiero nuevo.

Ana Liza: Ofrecen apartamentos de una, dos o tres alcobas. Y, los apartamentos que anuncian están clasificados según la zona de la ciudad donde están ubicados. ¿En qué zona lo quieres tú?

Perla Madonna: Aún no lo he decidido. Lo que sí sé, es que debe tener una buena vista, es decir, que cuando me asome a la ventana pueda ver algún paisaje y no, por ejemplo, otro edificio.

Ana Liza: Debemos, entonces, hacer una lista de los aspectos que son importantes para tí, de modo que los tengamos presentes al tomar la decisión.

Perla Madonna: Es una buena idea. Comencemos. Anota: precio, número de habitaciones, facilidades de pago,...

Ana Liza: Perla, me tengo que ir a clase. Pero, tan pronto como vuelva acabamos de hacer la lista y comenzamos a conseguir información sobre los apartamentos anunciados, para poder tomar una decisión.

Perla Madonna: Está bien, Ana. Te espero.

a. Determine exactamente cuál es la población de estudio en este caso. Además, diga cuál es la muestra que se está tomando y qué tamaño tiene.

b. Haga una lista de todos las características importantes, para la decisión que deben tomar Ana Liza y su hermana.

c. En el caso que nos ocupa, una de las características relevantes hace referencia al número de habitaciones de cada apartamento y otra, a la ciu-

dad donde está construido. ¿Encuentra usted alguna diferencia **esencial** entre los dos tipos de características mencionadas, en cuanto a la variación de ellas de elemento a elemento en la población? ¿Cuál? ¿Puede afirmarse que el número de habitaciones varía de un apartamento a otro? ¿Tiene sentido, en este caso, afirmar que la ciudad donde está construido el apartamento varía? Explique sus respuestas.

d. Con base en la respuesta que dio a la pregunta anterior, encuentre un adjetivo apropiado para calificar cada una de las características: número de habitaciones y ciudad donde está construido el apartamento.

e. Considere las características: barrio donde está ubicado el apartamento y calidad de los acabados del apartamento. ¿De qué manera se expresan los posibles resultados en cada caso? ¿Encuentra usted diferencias entre el **tipo** de valores que asume cada una de tales características? ¿Cuáles?

f. Considere las características: barrio donde está ubicado el apartamento y número de habitaciones que tiene. ¿Encuentra usted diferencias entre el **tipo** de valores que asume cada una de esas características? ¿Cuáles?

g. Considere las características: número de habitaciones y precio del apartamento. ¿Encuentra usted diferencias entre el **tipo** de valores que asume cada una de tales características? ¿Cuáles?

h. Con base en las respuestas que dio a las preguntas anteriores, clasifique las características que se presentan a continuación y diga qué criterio empleó para esa clasificación:

- Zona de ubicación dentro de la ciudad
- Area
- Calidad de los acabados
- Número de habitaciones
- Precio
- Piso
- Barrio

Formalicemos un poco

El concepto de variable está ligado estrechamente con las características, rasgos o atributos comunes que tienen los elementos de la población de estudio y, también está ligado con la variación que se da en dichos elementos con respecto a esos atributos.

Expliquemos lo dicho en el párrafo anterior, empleando la situación que se plantea en el diálogo. La población de estudio es el conjunto de todos los edificios nuevos de Bogotá, que son ofrecidos en venta en los avisos clasificados del periódico *El Desinformador*, durante un determinado lapso; y la muestra a la cual se hace referencia está constituida por los elementos de la población, es decir, los edificios que se ofrecen en venta el día que se sostiene el diálogo. Pues bien, todos los elementos de la muestra de estudio, al igual que los de la población de estudio, tienen muchos atributos: todos los apartamentos están ubicados en alguna zona de la ciudad, todos tienen área, todos tienen un cierto número de habitaciones, etc. y se podría hacer una lista extensa de atributos o rasgos que, considerados en conjunto, son los que hacen la esencia de los entes. A pesar de que todos y cada uno de los elementos de la población tienen los mismos atributos, de elemento a elemento hay variación en los valores que asume el atributo, y entonces se dice que el *atributo es variable*. Volviendo al ejemplo del diálogo, uno de los apartamentos puede estar ubicado en la zona 1, otro en la zona 3, etc.; uno puede tener 2 habitaciones y otro, 1 habitación, etc.

> Al hablar en estadística de una *variable* nos referimos a un atributo observable, —en los elementos de una muestra o de una población de estudio— que no asume el mismo valor para todos los elementos, es decir, toma dos o más valores.

En caso de que el atributo considerado tome el mismo valor para todos los elementos observados, en realidad no se trata de una variable sino de un atributo *constante*. En el ejemplo que estamos considerando, todos los apartamentos son nuevos y todos están situados en Bogotá, por tanto ni ciudad de ubicación, ni estado del apartamento son atributos variables.

Puesto que los atributos tienen diferente naturaleza, esto debe reflejarse en la forma de medirlos. Para efectos de producir una medida correspondiente a cada uno de los edificios observados en la muestra, no es lo mismo considerar el barrio en donde está ubicado que el área de construcción que tiene: en el

primer caso, la "medida" es una categoría, más exactamente un nombre; en cambio en el segundo caso, la medida es un número. El primer atributo describe a cada edificio cualitativamente, —en realidad, esa variable no produce medidas en el sentido usual de tal palabra; lo que produce es una clasificación— en tanto que el segundo lo describe cuantitativamente. El hecho descrito anteriormente da un criterio de clasificación de las variables: una variable puede ser *categórica* o *cuantitativa*.

Variables como el lugar de ubicación de una construcción, el sexo de una persona, la nacionalidad, la universidad donde estudia una persona, la calidad de una obra, el estado de un objeto que se va a comprar, el semestre que cursa un estudiante, etc. son variables categóricas. Variables como el área, el volumen, la edad, la estatura, el tiempo de duración de un suceso, la calificación obtenida en un examen, el número de hijos de una familia, el número de objetos defectuosos, el número de semestres cursados en la universidad, etc. son variables cuantitativas.

> Una variable se dice que es *categórica* si sus posibles valores son categorías de clasificación.
> Una variable se dice que es *cuantitativa* si los resultados que puede asumir son los resultados de medidas numéricas.

Con respecto al problema de los apartamentos, consideremos las dos variables barrio y calidad de terminados. Ambas son categóricas, sin embargo, existe una diferencia entre el tipo de categorías que establece cada una de esas variables. En el primer caso, se asignan nombres a los diferentes valores que puede tomar la variable. En el segundo caso, los valores que puede tomar la variable son algo más que nombres: son categorías que conllevan un juicio de valor que exige comparar a los diferentes elementos de la muestra con respecto a la variable en cuestión, para terminar ordenándolos. Variables categóricas del mismo tipo que barrio donde está ubicado un apartamento, se llaman *nominales*. Variables categóricas del mismo tipo que calidad de terminados de un apartamento, se llaman *ordinales*.

> Una variable categórica se llama *nominal* si los valores que puede asumir clasifican los elementos observados, pero no los ordenan. En caso de que los valores que pueda asumir la variable categórica, clasifiquen y ordenen los elementos observados, entonces se dice que la variable es *ordinal*.

Con respecto al problema de los apartamentos, consideremos las dos variables número de habitaciones y área. Ambas son cuantitativas, sin embargo, existe una diferencia entre ellas. En el primer caso, los valores que puede asumir la variable son números enteros: por ejemplo, puede ser que un apartamento tenga dos habitaciones y otro tenga tres, pero con certeza se sabe que no existe apartamento alguno que tenga un número de habitaciones que esté entre dos y tres. Para variables como ésta, en caso de que se representen sobre una recta todos los posibles valores que asume, la apariencia de la gráfica será una serie de puntos separados unos de otros; la separación entre los puntos puede ser eventualmente mayor o menor, pero existe esa separación. En el segundo caso, los valores que puede tomar la variable —por lo menos teóricamente— son **todos** aquellos que están en un intervalo determinado. Si se quiere representar sobre una recta los valores que asume una variable de este tipo, la apariencia de la gráfica será un segmento de recta, una semirrecta o una recta. Variables cuantitativas del mismo tipo que número de habitaciones de un apartamento, se llaman *discretas*. Variables cuantitativas del mismo tipo que área de un apartamento, se llaman *continuas*.

> Una variable cuantitativa se llama discreta si los valores que puede asumir están separados entre sí por una cierta cantidad. En caso de que los valores que pueda asumir la variable cuantitativa, sean todos los de un intervalo, entonces se dice que la variable es *continua.*

Para terminar, queremos hacer notar que el conjunto de todos los valores, que con respecto a una cierta variable, pueden tomar los elementos de una población de estudio es lo que en el capítulo anterior se definió como población de datos.

Algunos ejercicios

1.- En la sección anterior, ¿cuántos y cuáles criterios se dieron para la clasificación de las variables? Haga un esquema que muestre tal clasificación.

2.- Dé dos o tres ejemplos de las distintas clases de variables mencionadas en la sección anterior. Explique su respuesta.

3.- Clasifique las siguientes variables:

- El peso de un adulto
- La altura de los edificios del centro de Bogotá, de más de cuatro pisos
- El número de carros que posee un colombiano cualquiera
- La profesión de un grupo de amigos
- El color de un edificio
- El número de días que llueve en un mes del año
- El ingreso familiar
- La edad de las mujeres de una fábrica
- El número de libros de las bibliotecas de la universidad donde usted estudia
- El tiempo de experiencia laboral de un trabajador

4.- Determine qué variables son importantes en la elección de la universidad donde una persona puede estudiar. Clasifíquelas.

5.- Tahuro está realizando un estudio sobre los casinos de Bogotá. Le interesa principalmente conocer la edad promedio y la proporción de hombres y mujeres que entran en estos salones de juego.

a. ¿Cuál es la población de estudio?

b. ¿Cuáles son las variables de interés y de qué tipo son?

6.- Determine en cada caso, si se trata o no de una variable. En caso de ser variable, clasifíquela.

- La edad mínima para poder votar por primera vez
- El número de llamadas telefónicas que se hacen de una determinada línea durante un mes
- El número de clases que recibe usted este semestre, los miércoles
- La máxima calificación que puede obtener un alumno en un parcial, calificado sobre 50
- La calificación que puede obtener un alumno en el parcial descrito en el caso anterior
- El número de hijos que tiene la familia Pérez González
- El número de hijos que tiene una familia colombiana cualquiera
- La calificación que dan las personas a la atención recibida en el restaurante X
- Las materias que recibe un estudiante de segundo semestre en la Universidad X

7.- Para mañana, Stadi Shka debe llevar un ejemplo de una situación que eventualmente pueda estudiarse desde el punto de vista de la estadística. Ella piensa que podría resultar divertido hacer un test adecuado para medir memoria visual de los estudiantes de la Universidad donde ella estudia. Suponga que se acepta esto como tema de estudio.

a. ¿Cuál es la población de estudio?

b. ¿Cuál es la variable que se está tratando de medir?

c. Determine tres variables importantes de tener en cuenta al seleccionar las muestras y diga de qué tipo son los valores que toman.

d. Describa cómo conseguiría una muestra aleatoria para este experimento.

8.- Sexo es una de las variables consideradas en una encuesta que se practicó a los estudiantes de la Universidad. Se convino que 1 representa hombre y 0 representa mujer. Dé su opinión sobre la siguiente afirmación:

> Esta variable es cuantitativa porque los valores que asume son números.

9.- Explique lo que para usted significa la siguiente frase:

> Toda variable cuantitativa se puede convertir en variable categórica.

¿Cree que esa afirmación sea verdadera? Explique su respuesta.

10.- Estudiantes de quinto semestre de Ciencia Política realizaron una investigación sobre el proceso de Elección Popular de Alcaldes en los municipios de la sabana de Bogotá. Tuvieron como referencia los casos particulares de los municipios de Chía y Cota en cuanto a las elecciones de alcalde de 1988 y 1990.

Al inicio de su investigación se plantearon unos objetivos: buscaban conocer hasta qué punto la Elección Popular de alcaldes constituyó una nueva vivencia política no sólo en el comportamiento de los partidos políticos sino también en la participación del electorado. Su investigación se realizó

en torno a una serie de variables: número de votantes de EPA en 1988 y en 1990; participación del electorado (número de personas que votaron con respecto al número de personas que están en capacidad de hacerlo); orientación partidista de los electores; orientación partidista de los candidatos; resultados electorales de 1988 en relación con los de 1990.

a. Defina claramente cuál es la población de estudio y señale cuál es la muestra con base en la cual se va a inferir.

b. Clasifique las variables mencionadas anteriormente.

c. ¿Qué otras variables consideraría usted para tal estudio?

d. ¿Cree usted que la participación de los floricultores en el proceso electoral sea un indicador apropiado para medir la compra de votos en la elección? ¿Por qué?

11.- Estudiantes de Ciencia Política quieren determinar si el mecanismo de la Consulta Popular empleado el año pasado (1990) por el partido liberal en las elecciones de marzo sirvió como mecanismo "modernizador" y "oxigenizador" del partido liberal, para así predecir si el empleo de este mecanismo por parte de otras fuerzas políticas podría contribuir de manera significativa a la modernización del régimen electoral colombiano. Ellos centrarán su estudio en los municipios cundinamarqueses que presentan características políticas (determinada tendencia partidista) similares a las de la capital y tomarán tres municipios: Chía, Tabio y Cota para observar el comportamiento de los ciudadanos frente a la Consulta Popular. Tomarán como variables: nivel de abstención en las elecciones de esa fecha; partidos políticos (liberal, conservador, etc.) potencial electoral y número de votantes frente al proceso.

a. Defina la población de estudio y la muestra.

b. Clasifique las variables mencionadas anteriormente.

c. ¿Qué otras variables tendría en cuenta usted para el estudio? ¿Por qué?

12.- La desinstitucionalización de los espacios de negociación es un problema que ha suscitado varios interrogantes. Estudiantes de sexto semestre de Ciencia Política desean analizar el problema. Para eso buscan verificar hasta qué punto las reformas de descentralización hechas en 1986 amplia-

ron espacios para el "acceso a la formalidad política" y permitieron de alguna manera la "reinstitucionalización de conflictos sociales". Los politógos se basaron en el análisis de algunas variables: partidos políticos, espacios nuevos de negociación, canales de participación ciudadana. Es importante señalar que harán su investigación basados en una muestra (teniendo en cuenta las elecciones de alcalde en 1990 en el municipio de Tabio) para inferir conclusiones acerca de la correspondiente situación en los municipios aledaños a Bogotá.

a. Mencione cuál es la población de estudio.

b. ¿Cuáles objetivos además de los planteados consideraría usted?

c. Clasifique las variables mencionadas y diga cuáles se escapan a la luz del marco conceptual.

13.- La adición de partículas diminutivas a las palabras es casi una costumbre en la mayoría de hispanoparlantes. Este fenómeno se ha visto en Colombia en regiones como Nariño y Boyacá principalmente.[13]

Un antropólogo observó en el municipio de Llano Blanco, cerca de Villa de Leyva (Boyacá), que había cierta tendencia a adicionar terminaciones diminutivas preferiblemente a palabras que se refieren a alimentos, ya que establecer mayor "amistad" con el alimento ocasiona que éste no haga tanto daño al ingerirse, es decir que no cause enfermedad. Para comprobar la hipótesis de que existe cierta relación entre el uso de diminutivos y las concepciones de enfermedad en esa comunidad campesina, Jorge Morales Gómez decidió adelantar una investigación en Llano Blanco durante el período comprendido entre 1981 y 1983. La investigación consideró tres aspectos: 1) la lexicografía, 2) las formas de alimentación y 3) los conceptos de salud y enfermedad, los cuales fueron observados mediante el registro de conversaciones cotidianas entre los habitantes del municipio.

Elementos relevantes considerados en el estudio fueron:

- Ubicación del municipio
- Actividad económica de los habitantes

13 Problema tomado de "El diminutivo y la noción de enfermedad", investigación realizada por Jorge Morales Gómez, antropólogo de la Universidad de los Andes. Tomado de *Estudios sobre español de América y lingüística afroamericana.* Bogotá: Instituto Caro y Cuervo, 1989, pp. 127-136.

- Forma generalizada de tenencia de la tierra
- Concepto de comunidad
- Clase social
- Clases de palabras a las que se les agrega diminutivos
- Forma generalizada de alimentación
- Alimentos básicos que se consumen
- Concepto de salud
- Concepto de enfermedad

a. ¿Cuál es el problema de estudio?

b. ¿Cuál es el objetivo de la investigación?

c. ¿Cuál es la hipótesis?

d. ¿Cuál es la población?

e. De los diez puntos mencionados, ¿cuáles son variables? ¿De qué tipo de variable se trata? Justifique su respuesta.

f. Mencione posibles categorías de esas variables.

¿Qué vamos a hacer y cómo lo vamos a hacer?

Introducción

En los capítulos anteriores se han realizado principalmente dos cosas: una, intentar motivar y justificar el estudio de la estadística en un curso de Ciencias Sociales, y otra, aproximarnos a los tres conceptos más elementales con que trabaja permanentemente la estadística.

Vamos a suponer que estamos frente a un problema social muy complejo, cuyas características exigen el empleo de la estadística como herramienta para solucionarlo. También vamos a suponer que ya están dados los primeros pasos requeridos en cualquier investigación, es decir, ya está bien definido el problema, está delimitada la población de estudio, están claramente determinados los objetivos de la investigación, ya están determinadas las variables que se quiere estudiar, y aún más, ya se cuenta con una muestra de datos. En este punto cabe preguntarnos ¿y, ahora qué? Pues bien, ahora que contamos con una muestra de datos debemos obtener de ellos toda la información que sea útil para lo que interesa en la investigación y esto se logra *organizando* y *resumiendo* la información de la muestra de datos. La *estadística descriptiva* se encarga precisamente de esa tarea: describir las muestras de datos.

Pero, aunque la tarea de describir es importante e imprescindible como parte del método estadístico, en la mayoría de los casos no es suficiente, es decir, debe hacerse algo más que conduzca a la solución del problema que está planteado en términos **muy generales**. En efecto, se requiere *interpretar* y *generalizar* los resultados obtenidos en la muestra de manera que se puedan aplicar las conclusiones obtenidas a la población de datos de donde provino la muestra de datos tomada. La *estadística inferencial* se encarga precisamente de esa tarea: emplear la información contenida en una muestra de datos para hacer predicciones sobre la correspondiente población de datos y además para justificar la toma de decisiones.

En los tres capítulos siguientes "Organización y resumen gráfico de datos", "Medidas de tendencia central" y "Medidas de dispersión" centraremos nuestra atención en herramientas proporcionadas por la estadística descriptiva. Para describir una muestra se cuenta con diferentes clases de medios, cada uno de los cuales tiene su propia importancia dentro del método estadístico y su importancia relativa con respecto a las demás herramientas. Las herramientas a las que nos referimos permiten organizar datos, (tablas, distribución de datos) representarlos gráficamente (diagramas) y describir (medidas de posición y medidas de dispersión) los valores que asume una variable en una muestra.

En este capítulo damos a conocer la metodología de trabajo que hemos adoptado para abordar el conocimiento de las distintas herramientas con las que trabajaremos en lo que queda del curso.

¿Por qué inventar herramientas?

(Algunos alumnos del curso están charlando mientras llega el profesor de matemáticas.)

Stadi Shka: Ahora sí comienza mi sufrimiento en la clase de matemáticas. Estoy muy asustada.

Estrella: ¿Por qué?

Stadi Shka: ¿No oíste al profesor? Vamos a comenzar a emplear herramientas de estadística para organizar, manejar e interpretar los datos que recojamos en el experimento que vamos a realizar. ¡Uy, qué susto! Ahora se me vienen encima una cantidad de fórmulas, de definiciones y con mi mala memoria ya no voy a saber qué fórmula usar ni cómo hacerlo; además, recuerda que soy malísima para los cálculos.

Ana Liza: ¡Ay, no seas bobita, Stadi Shka! Tú no atendiste a todo lo que dijo el profesor. Eso que te está asustando tanto, fue lo último que él dijo. Pero lo que él se propone es que en clase logremos que se desarrolle el proceso de construcción de cada una de las herramientas que vamos a usar; por tanto, no creo que el empleo de ellas sea difícil.

Estrella: Sí, pero de todas maneras, las demostraciones o deducciones suelen ser muy complicadas.

Tahuro: Pero... si no me equivoco, la idea que planteaba el profesor es la de re-inventar las herramientas; eso nos remontaría al origen de ellas; es decir, ese proceso debería aclararnos más las ideas que lo que lo podría hacer la sola deducción de una fórmula.

Ana Liza: Sí. Creo que lo que más le interesa al profesor es que entendamos que cada herramienta tiene su justificación (es decir, que no fue inventada por capricho de alguien, sino que surge a partir de necesidades concretas) y que conocer esa justificación sirve para aplicar racionalmente las fórmulas a las que se llegue.

Askanio: Y, ¿cómo sería un ejemplo del proceso al que hacía referencia Tahuro? Intentemos construir una herramienta que pueda ser útil en estadística.

Ana Liza: A ver... No se me ocurre... ¿Cómo re-inventar el promedio?

(*En ese momento llega el profesor, un poco retardado, y Ana Liza le pide que les dé un ejemplo adecuado de la forma como se espera que ellos trabajen.*)

Profesor: Bien. Veamos... Supongan que un delegado de la Junta de Consumidores recoge algunos precios de venta de un determinado artículo, en varios almacenes del sector norte de la ciudad y encuentra la siguiente información: $350, $380, $300, $385, $315, $367, $365, $380, $310, $385. Si el delegado quiere obtener, a partir de los datos que recogió, un número que sirva para representar el precio de costo de tal artículo, ¿cuál creen ustedes que pueda ser tal número?

Askanio: Pues, eso es fácil. Hagamos el promedio de esos diez datos.

Stadi Shka: ¡Un momento, por favor! Yo habría tomado otro número como representante. No estoy segura si es correcto, pero creo que yo habría escogido el menor de los diez números, es decir habría escogido a $300 como el número más próximo al precio de costo del artículo que se está mencionando.

Askanio: Verdad, ¿no? Es más adecuado tomar $300 que lo que resulte al hacer el promedio pues... con toda seguridad todos los almacenes visitados le ganan algo al artículo: posiblemente, unos ganen más que otros pero lo que es seguro es que el precio de costo del artículo debe ser menor de $300.

Estrella: Y, claro eso no significaría que estemos seguros de que el precio de costo del artículo sea $300, sólo que $300 es un número cercano, pero superior al precio de costo. Además, creo que tendríamos que aclarar que ese número es un buen representante de lo que se quiere sólo si los diez datos recolectados fueron adecuadamente tomados, es decir, si son representativos de lo que en ese aspecto sucede en el sector norte de la ciudad.

Ana Liza: Creo que Stadi Shka, Askanio y Estrella se sobraron en la respuesta. ¿No es cierto?

Profesor: De acuerdo. Ahora, supongan que el administrador de un edificio en el cual viven 50 familias tiene que alquilar un número determinado de sillas para colocarlas en el salón comunal para una reunión trimestral que debe efectuarse con los propietarios de los apartamentos. El administrador sabe que en las diez últimas reuniones del mismo tipo han asistido 23, 18, 25, 25, 30, 20, 28, 22, 33, 17 personas. Si el administrador debe decidir cuántas sillas alquilar, con base en la información que tiene, ¿cuál creen ustedes que sea el número más representativo en este caso?

Tahuro: Si yo fuera el administrador, pensaría en dos aspectos para tomar la decisión: primero, desearía que todos los asistentes pudieran sentarse; segundo, desearía no alquilar en vano unas cuantas sillas. Por tanto, creo que yo alquilaría unas 33 sillas, lo que representa el número más grande de asistentes que ha habido en ese tipo de reuniones a partir de las últimas diez.

Ana Liza: En este caso, también se habría podido escoger el promedio, ¿no es cierto?

Profesor: Ciertamente. Sin embargo, creo que la respuesta de Tahuro para este caso es más adecuada. En cambio, para el primer caso que consideramos es evidente que el promedio definitivamente no sirve.

Askanio: Entiendo lo que se ha dicho hasta el momento, pero... ¿en qué quedó el proceso de re-inventarnos las herramientas? Hasta ahora no veo que hayamos inventado ninguna.

Profesor: Y, ¿los demás qué opinan?

(*Chepa, quien había permanecido en silencio, habló.*)

Chepa: Yo creo que hemos encontrado dos herramientas estadísticas cada una de las cuales es apropiada para una de las dos situaciones planteadas.

Profesor: ¡Correcto, Chepa! Y, ¿qué nombre sugieres que le demos a cada una de tales herramientas?

Chepa: A la primera que encontramos, *mínimo* y a la segunda, *máximo*.

Profesor: Sí; esa respuesta es correcta.

Askanio: Y, ¿así de fácil?

Profesor: Bueno, aquí debemos hacer aclaraciones sobre algunas condiciones que deben cumplir los objetos que estamos re-inventando. En primer lugar, al igual que toda herramienta, las herramientas estadísticas son *instrumentos* que deben servir a quien las emplea para facilitar un cierto trabajo. Es, entonces, esencial saber claramente en qué consiste el trabajo que queremos realizar para poder encaminar inteligentemente nuestros pasos en la búsqueda del instrumento. Es decir, es imprescindible saber *qué vamos a hacer* y para qué lo vamos a hacer aunque no conozcamos aún cómo se realizará. En los dos ejemplos que trabajamos se quería obtener un número que representara un conjunto determinado de datos y para elegir ese representante se tuvo en cuenta además de la composición específica del conjunto, el propósito que tenía la elección de dicho representante.

(*Tahuro, insistiendo.*)

Tahuro: Sí, pero uno no se espera que una herramienta sea tan simple.

Profesor: Relacionados con ese punto hay dos aspectos que me gustaría mencionar. En primer lugar, la complejidad no es un requisito para la construcción de una herramienta. La complejidad de una herramienta está estrechamente ligada a la naturaleza del trabajo para el cual se ha construido. Por tanto, es perfectamente justificado que existan herramientas muy simples, y también herramientas muy complicadas. En segundo lugar, quiero resaltar una condición que ha estado presente en las respuestas que ustedes han dado a las preguntas que yo he formulado: la intuición. La intuición es una facultad que encamina, por lo menos, las primeras actividades del supuesto inventor o descubridor hacia el objetivo final. Es posible que la intuición en algunos casos se quede corta para lograr la construcción de una herramienta sofisticada; de acuerdo. Pero lo que no debería suceder es que una buena herramienta, pro-

duzca un resultado contrario al que daría una buena intuición. De manera que si ustedes se aproximan intuitiva y correctamente a una posible solución de un problema, casi con seguridad tendrán éxito en hallarla; y no se sorprendan si en ocasiones esa solución es muy sencilla. Además, tengan presente que buscamos una herramienta que sea *universal*: esto es, una herramienta que sea útil para la mayor parte de los casos en los que se quiera encontrar un único número que sea *representativo* de un conjunto de datos.

Stadi Shka: Lo que ustedes han dicho me tranquiliza bastante. ¡No lograba imaginarme cómo podría yo llegar a inventarme una fórmula o cómo podría encontrar la deducción de otra!

Askanio: Propongo que se haga un resumen de las conclusiones a las que hemos llegado.

Stadi Shka y Tahuro: (*Al tiempo.*) De acuerdo.

Estrella: Que lo diga Ana Liza.

Ana Liza: En resumen, hemos visto que una herramienta estadística debe tener ciertas características: finalidad, aplicabilidad, sencillez (si es posible), "universalidad", compatibilidad con una buena intuición. Y, en general, las herramientas que se van a redescubrir en este curso no surgen gratuitamente sino que se llega a ellas a través de un proceso —corto o largo, sencillo o complejo— basado en la necesidad que presentan las situaciones concretas que nos interesan.

Profesor: Muchachos, se nos acabó el tiempo. Nos vemos mañana.

Organización y resumen gráfico de datos

Introducción

En este capítulo se estudiarán algunas de las herramientas que permiten organizar y resumir gráficamente la información que se considere pertinente obtener a partir de una muestra.

La importancia de tratar estos temas radica en que siempre que se aborde un problema desde el punto de vista de la estadística —con el fin de llegar a obtener generalizaciones y a hacer inferencias sobre la población—será necesario tomar como base muestras de la población de estudio y describirlas. Los primeros pasos del proceso son, precisamente, la organización de los datos y el resumen de los mismos. El resumen de los datos puede hacerse gráfica o numéricamente. En este capítulo nos interesa hacerlo gráficamente.

Por lo general, antes de hacer la representación gráfica de un conjunto de datos conviene organizarlos en forma tabular con el fin de realizar el proceso de la manera más eficiente posible. Es esa la razón por la cual comenzaremos con una breve sección sobre tablas.

Tablas

Con el objeto de decidir si es necesario o no prestar un servicio de transporte a una cierta comunidad universitaria, se quiere determinar cuál es el medio de transporte más utilizado por los estudiantes de la universidad para llegar a clase diariamente. Se hizo una encuesta a algunos estudiantes, seleccionados al azar, y los resultados obtenidos son:

buseta	bus	a pie	a pie	a pie
bus	buseta	moto	a pie	moto
a pie	buseta	a pie	auto	buseta
buseta	moto	buseta	a pie	bus
buseta	moto	a pie	buseta	buseta
a pie	auto	auto	buseta	bus
buseta	buseta	bus	bus	moto

a. Dé una ojeada rápida a los resultados obtenidos a partir de la encuesta y diga:

- Cuántas personas fueron encuestadas.
- Cuántos y cuáles medios de transporte son empleados por ellas.
- Cuál es el medio de transporte más empleado por ellas.

Después de responder las preguntas anteriores, usted estará de acuerdo en que la forma que se utilizó para presentar la información no es la más adecuada si se quiere que quien la mire obtenga rápidamente una cierta cantidad de información. La tarea es, entonces, encontrar una mejor manera de presentar dicha información y seguramente usted ya tiene sugerencias al respecto.

b. Diga cuál es la variable que se está midiendo y de qué tipo es. ¿Cuántos y cuáles valores asume dicha variable?

c. Complete la tabla que se da a continuación. (Utilice la segunda columna con el fin de hacer eficiente el proceso de contar.)

Valores	**Conteo (marcas)**	**Total de marcas**

Si usted construye una tabla, empleando la primera y última columnas de la tabla anterior obtiene lo que se llama una *tabla de distribución de frecuencias*.

> Una *tabla de distribución de frecuencias* contiene dos columnas; una de ellas, la primera, muestra todos los posibles valores que asume la variable y la segunda, para cada valor de la variable, muestra el número de veces que se presenta dicho valor, en el contexto en el que se está trabajando; tal número se denomina *frecuencia*.

d. Dé una ojeada rápida a la tabla de distribución de frecuencias correspondiente al problema que estamos tratando y responda las tres preguntas planteadas inicialmente.

Si se observan cuidadosamente las dos maneras de presentar la información que se han mostrado en esta sección, se ve que la diferencia esencial radica en la organización: en la clasificación que se hace en el segundo caso y no en el primero; y eso incide en la **cantidad** de información que, en cada caso, se brinda en el **menor tiempo posible**.

Diagramas

(*Muy cerca de la universidad hay por lo menos tres cafeterías, a las cuales vamos a llamar A, B y C. Chepa y Estrella están tomando café en una de ellas.*)

Chepa: Estrella, ¿acabaste de trabajar el problema que te asignó el profesor?

Estrella: Todavía no he terminado; ya conseguí la información relacionada con el problema, pero me falta representarla gráficamente. Y, es ahí donde tengo dudas sobre cómo hacerlo. Ayúdame, ¿quieres?

Chepa: Yo también tengo mis dudas. Pero, tratemos de aclararlas.

Estrella: El enunciado del problema afirma que durante los días de clase, en la cafetería A almuerzan más estudiantes de la universidad que los que almuerzan en la cafetería B. Mi trabajo consiste en tomar una muestra de estudiantes de la universidad, con el objeto de determinar si, para dicha muestra, la afir-

mación es verdadera o no lo es; y debo presentar una gráfica en la que se haga evidente la información que obtenga.

Chepa: Pues, tu problema y el mío son muy similares. Cuéntame qué has hecho para responderlo.

Estrella: Tomé una muestra aleatoria de 100 estudiantes de la universidad y a cada uno de ellos les pregunté cuál es el sitio donde almuerzan con más frecuencia durante los días de clase. Y, después de clasificar las respuestas que me dieron, construí la tabla de distribución de frecuencias. (*Estrella saca su cuaderno y muestra la siguiente tabla*)

Lugar	**Frecuencia**
Cafetería A	16
Cafetería B	20
Cafetería C	24
Casa	20
Otros	20

Chepa: Entonces, en esa muestra **no** es cierto que haya más estudiantes de la universidad que almuercen en la cafetería A que en la B. Y, si ese es uno de los mensajes que debe dar la gráfica, entonces, creo que podría ser del siguiente estilo. (*Chepa dibuja.* La Gráfica 1 se presenta en la página siguiente.)

Estrella: ¡Claro! En esa gráfica es muy fácil comparar entre sí las frecuencias de las diferentes categorías en que se clasificó la variable.

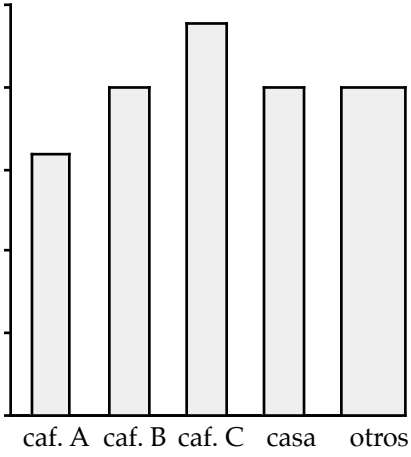
(*Llega Askanio*)

Askanio: ¡Hola amigas! ¿Qué hacen?

Chepa y Estrella: (*en coro*) Salsa de tomate. ¿No ves?

(*Askanio mira el diagrama que hizo Chepa, es decir, la Gráfica 1 y se refiere a él*)

Gráfica 1



Askanio: ¿Qué representa ese dibujo?

Chepa: ¡Ay, Askanio! Tú y tus preguntas... ¿Es que acaso no ves?

Askanio: ¡Perdón! ¡Qué genio!

(*Askanio se va*)

Chepa: Ahora, ayúdame a responder el problema que me asignó el profesor. El enunciado afirma que durante los días de clase, la mayoría de los estudiantes de la universidad almuerzan en la cafetería C. Al igual que tú, debo tomar una muestra de estudiantes de la universidad y debo presentar una gráfica que permita ver muy claramente si la afirmación se cumple o no, para la muestra.

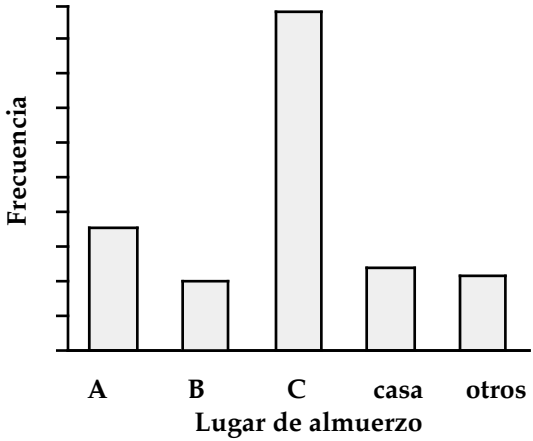
Estrella: Tu problema es idéntico al mío.

Chepa: Sí, creo que sí. Yo tomé una muestra de 100 estudiantes de la universidad, les pregunté su preferencia al respecto, y aquí tengo los resultados.

Estrella: Pues hagamos el dibujo de la misma manera que el anterior.

(*Después de un momento, tienen la Gráfica 2, que se presenta en la página siguiente*)

Gráfica 2



(*Regresa Askanio, acompañado de Ana Liza. Se sientan y piden un café.*)

Ana Liza: ¡Hola, muchachas! ¿Cómo les va?

Askanio: ¿Ya les pasó el mal genio?

Estrella: ¡Discúlpanos! Estábamos muy ocupadas, pero ya terminamos nuestra tarea.

(*Askanio ojea el último diagrama*)

Askanio: De manera que... la mayoría de las personas, a quienes se refiere ese diagrama, almuerzan en el sitio que ustedes llaman C.

Ana Liza: Mirando la gráfica rápidamente puede parecer que la mayoría de las personas, a las que se refiere, almuerzan en la cafetería C, pero yo no estaría tan segura. Puede ser que sí, como puede ser que no. Eso depende; para saberlo con certeza, tendríamos que sumar las frecuencias de todas las categorías, sin incluir la correspondiente a C, y comparar ese número con la frecuencia de C. Sólo, así podríamos saber si lo que dice Askanio es cierto o no.

(*Chepa mira la tabla*)

Chepa: Ana Liza tiene razón: de los 100 estudiantes que entrevisté, 49 prefieren almorzar en la cafetería C y los demás, 51, prefieren no almorzar allí.

Estrella: ¡Fíjate, Chepa! Por muy poco, la afirmación que hizo Askanio –que es la misma hecha en el enunciado de tu problema– no es cierta. Sin embargo, eso no se hizo completamente evidente en la gráfica. Por tanto,...

Chepa: ¡Ya me di cuenta, Estrella! Para este caso, este tipo de gráfica **no** es la más conveniente. Y, entonces ¿qué hago?

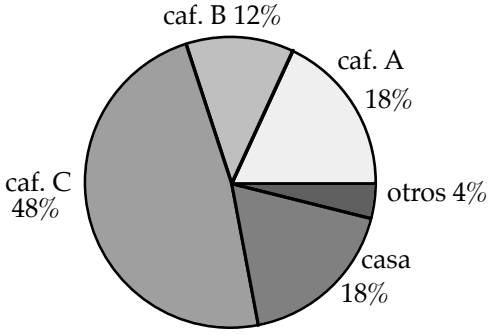
Askanio: ¿Qué es lo que pasa?

Chepa: Mi problema es el siguiente: tengo una variable categórica que asume varios valores y quiero encontrar un tipo de gráfica que permita comparar muy fácilmente la frecuencia de una de las categorías con la frecuencia de las demás categorías, o con el total. Dicho en otras palabras, la gráfica que haga debe permitir una comparación eficiente de una parte con el todo.

Ana Liza: ¡Ah! Para eso puedes hacer una "torta". Hagámosla.

(*Ana Liza saca de su bolso un transportador, hace cálculos y dibuja. Después de algún tiempo muestra el siguiente diagrama.*)

Gráfica 3



Distribución de la variable lugar de almuerzo

Estrella: Este diagrama sí dice que quienes almuerzan en la cafetería C **no** son la mayoría. Hay más del 50% de las personas encuestadas que no almuerzan en tal lugar.

Chepa: ¡Listo! Gracias muchachos. Voy a pasar a limpio mi tarea. Adiós.

a. ¿Cuál es la variable que se está midiendo? ¿De qué tipo es? ¿Cuántos y cuáles valores asume la variable?

b. Observe detenidamente el primer diagrama que se presentó en el diálogo, (Gráfica 1) ¿tiene Askanio toda la información necesaria para entender e interpretar la información que se pretende dar? ♣*¡O sea, que esas viejas histéricas regañaron a Askanio injustamente*♣ *Explique su respuesta.*

c. Si el diagrama al que se hace referencia en el item anterior tuviera toda la información necesaria para poder leerlo e interpretarlo, ¿cómo se sabría cuántas observaciones hay en cada clase?

d. Considere la muestra seleccionada por Estrella. Con base en la correspondiente tabla de distribución de frecuencias, compare las frecuencias de las categorías *Cafetería B* y *Otros*. Haga la misma comparación con base en el diagrama (Gráfica 1). ¿Deberían coincidir las dos respuestas anteriores? ¿Coinciden? (¡Mire bien!)

e. Suponga que Chiripa y Hazard tomaron también dos muestras de estudiantes de la universidad, les preguntaron en dónde almuerzan frecuentemente los días de clase y quieren comparar los resultados obtenidos en las dos muestras. A continuación se presenta la tabla de distribución de frecuencias:

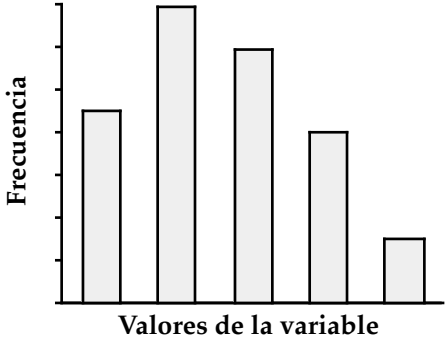
Lugar	**Frecuencia**	
	Chiripa	**Hazard**
Cafetería A	24	36
Cafetería B	10	15
Cafetería C	14	21
Casa	12	18
Otros	20	30

Represente gráficamente los resultados obtenidos para cada una de las dos muestras. Compare el aspecto de las dos gráficas y exprese esa comparación en palabras. ¿Cuál es el tamaño de cada una de las dos muestras? ¿Qué porcentaje de los encuestados por Hazard prefieren la cafetería C? ¿Qué porcentaje de los encuestados por Chiripa prefieren la

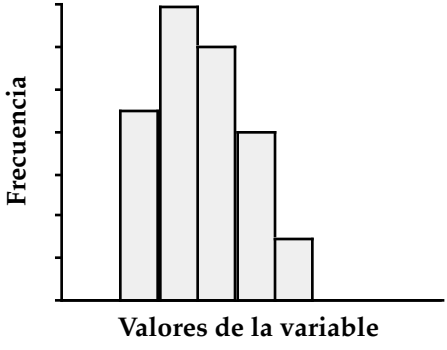
cafetería C? ¿Reflejan las gráficas que usted hizo el hecho de que esas proporciones son iguales? ¿Qué cambio tendría que hacer en sus gráficas para que al compararlas sea evidente la comparación? Ahora sí, vuelva a presentar las gráficas que le permitan comparar fácilmente los resultados obtenidos en las dos muestras.

f. ¿Existe alguna diferencia esencial entre los diagramas que se muestran a continuación? ¿Cuál es? ¿Refleja esa diferencia entre los diagramas alguna diferencia entre las variables cuya distribución representan? Explique su respuesta. (Para dar su respuesta, tenga en cuenta si cree que alguna de las dos gráficas es más adecuada para representar un determinado tipo de variable.)

Gráfica 4



Gráfica 5



g. ¿Cree usted que el diagrama empleado en el caso que trabajó Estrella (Gráfica 1) es adecuado? ¿Por qué? Explique claramente por qué el mismo tipo de diagrama no resultó adecuado para la situación que trabajó Chepa (Gráfica 2).

h. Explique cómo cree que se hace un diagrama de "torta" y diga en qué casos es conveniente usarlo.

i. Haga un resumen ilado de las conclusiones a las que llegó después de haber leído el diálogo y después de haber respondido las preguntas anteriores.

Formalicemos un poco

En esta sección vamos a tratar de hacer explícitas algunas consideraciones que conviene tener en cuenta cuando se quiere representar gráficamente la información contenida en una muestra.

Para organizar y resumir la información contenida en una muestra existen técnicas –herramientas– más o menos sencillas de usar. Resulta interesante conocerlas y saberlas emplear. Sin embargo, no hay que olvidar que la función de esas técnicas es contribuir al manejo eficiente de la información. Al hacer una gráfica, a partir de una muestra, lo que se pretende es que quien la observe se haga rápida y fácilmente una idea aproximada de lo que expresan los datos.

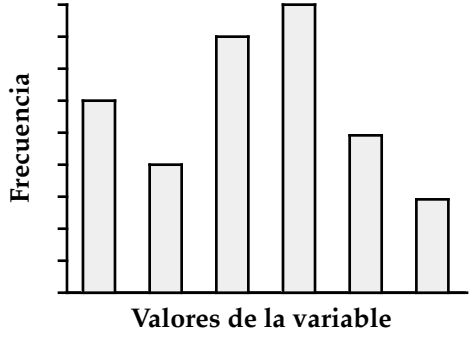
Puede pensarse, entonces, que existen muchas reglas que rigen el empleo de las gráficas. En realidad no es así; lo que hay son razones que justifican el empleo de un determinado tipo de gráfica en determinado caso, y esas razones se refieren básicamente a tres aspectos. Uno de ellos es el tipo de variable que se quiere representar; por ejemplo, como hay diferencias esenciales entre las variables categóricas y las variables cuantitativas, es natural que esas diferencias se reflejen en las correspondientes gráficas. El otro aspecto, está relacionado con **qué** se quiere presentar en la gráfica y para qué; por ejemplo, si se quieren comparar dos gráficas referentes al mismo asunto es conveniente usar la misma escala para que la comparación sea evidente. Y, el tercer aspecto se refiere a cuestiones elementales de presentación como son dar a conocer el título de la gráfica y hacer explícitos todos aquellos detalles que permiten a cualquier persona entender de qué se trata la gráfica.

El siguiente ejemplo intenta aclarar lo dicho anteriormente. Si se quiere mostrar gráficamente cómo se distribuye una variable categórica, se puede construir un *diagrama de bloques*. En tal caso, como los valores que asume la variable no son numéricos no interesa dar un orden a la disposición de los mismos y por consiguiente no se puede hablar de la clase X que está entre las clases Y y Z, por lo cual los rectángulos que las representan se hacen separados unos de otros. En cambio, si se quiere mostrar la distribución de una variable cuantitativa continua no conviene hacerlo con un diagrama de bloques pues si se hiciera, se daría la impresión de que la variable no toma todos los valores que realmente puede tomar. Pero, ¿qué es un diagrama de bloques?

> Un *diagrama de bloques* es una gráfica que se emplea para representar la distribución de una variable categórica. Consta de una serie de rectángulos, cada uno de los cuales representa una categoría de la variable. Las bases de los rectángulos están sobre una misma recta y se nombran con los valores que toma la variable cuya distribución se quiere representar. Las bases de todos los rectángulos tienen la misma longitud y la altura de cada uno de ellos es proporcional al número de observaciones de la muestra que están incluidas en cada clase. Los rectángulos que conforman la gráfica están separados entre sí para indicar que entre uno y otro valor de la variable no hay más valores.

El diagrama de bloques es muy fácil de construir y muy útil siempre que se comparen entre sí las frecuencias de las diferentes categorías en que se clasifica la variable, pues esa comparación se establece atendiendo al rectángulo que tenga mayor o menor altura. Un diagrama de bloques tiene la siguiente forma:

Gráfica 6



Sin embargo, si se quiere comparar la frecuencia de una de las clases con la frecuencia de las demás clases o con el total de observaciones que constituyen la muestra, no es el diagrama de bloques el más adecuado. Para esos casos, es mejor construir un *diagrama circular*.

> Un *diagrama circular* es una gráfica que se emplea para representar la distribución de una variable categórica. Para construirlo se utiliza un círculo: se divide en tantos sectores como categorías tenga la variable. El tamaño de cada sector (o sea del ángulo central correspondiente) debe ser proporcional al número de observaciones de la muestra que están incluidas en cada clase.

Veamos en un ejemplo cómo se construye un diagrama circular para representar una variable categórica. Suponga que tiene una muestra de 50 estudiantes de la Universidad, se les pregunta en qué semestre van y lo que interesa es clasificar las respuestas obtenidas en una de tres clases, definidas así:

- De primero a tercer semestre: clase A
- De cuarto a sexto semestre: clase B
- De séptimo en adelante: clase C

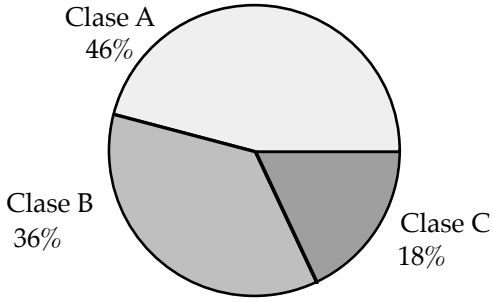
A continuación se da la distribución de frecuencias que se obtuvo:

Semestre	**Frecuencia**
Clase A: de primero a tercero	23
Clase B: de cuarto a sexto	18
Clase C: de séptimo en adelante	9

En este caso debe dividirse el círculo en tres sectores, cada uno de los cuales representará una de las clases. El tamaño de cada sector deberá ser proporcional a la frecuencia de la clase que representa. Por tanto para calcular el valor del ángulo bastará hacer una regla de tres. Por ejemplo, el ángulo central correspondiente a la clase A deberá medir 165,5 grados.

De manera similar se obtiene el tamaño de los otros sectores. Y con esa información se construye el diagrama circular. A continuación se presenta:

Gráfica 7



Distribución de la variable semestre

Otras gráficas

(*Hazard llega a la casa de Chepa.*)

Hazard: ¡Estrella y tú se lucieron en clase!

Chepa: Gracias, Hazard. Y, tú, ¿cuándo vas a exponer?

Hazard: Mañana. Mi exposición tiene que ver con un problema real en el que estoy metido. ¿Me quieres ayudar?

Chepa: ¡Claro!

Hazard: Mi hermana, la que estudió pre-escolar, quiere montar un jardín infantil en el barrio donde vivimos, pero no tiene el dinero necesario y está buscando que nuestro papá la financie. Para lograrlo debo convencer a papá de que la idea es buena, pues muy seguramente habrá una gran demanda.

Chepa: ¡Uy! es un problema complejísimo. ¡Hay una gran cantidad de elementos involucrados y muchas interrelaciones entre ellos!

Hazard: Sí. Ya tengo construido, más o menos satisfactoriamente, el modelo del problema social correspondiente. Creo que una de las variables de interés es el número de hijos que tiene cada familia residente en el barrio; y otra, es la edad de los hijos de las familias.

Chepa: Y, ¿en qué te puedo ayudar?

Hazard: Mira: tomé una muestra aleatoria de 50 familias, visité a los dueños de casa y les hice responder 10 preguntas relacionadas con los aspectos que determiné como más importantes para lo que me interesa. Como mi objetivo principal es convencer a papá de que la idea de mi hermana es buena, quiero presentarle, de la manera más adecuada, la información que recogí, y es ahí donde puedes ayudarme.

Chepa: Pues, a comenzar. Muestra las encuestas.

(*Hazard busca las encuestas y las da a Chepa.*)

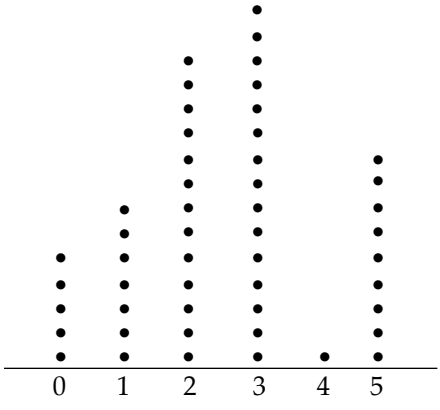
Hazard: Comencemos por organizar los datos sobre el número de hijos; los encuentras como respuesta a la tercera pregunta. Yo ya estuve mirando las respuestas que dieron a esa pregunta y sé que los valores que toma la variable son: 0, 1, 2, 3, y 5.

Chepa: Comienzo a dictarte las respuestas a la pregunta 3: 0, 3, 5, 1, 1, 2, 2, 2, 2, 3, 1, 2, 2, 0, 5,...

(*Chepa dicta los 50 valores.*)

Hazard: ¡Listo! Mira un bosquejo de la forma que tendrá la gráfica cuando se haga elegantemente.

Chepa: ¡Qué maravilla! Ese método de presentar la información es mejor que dar una tabla de distribución de frecuencias; pues a la vez que organiza la información, da una representación gráfica de la misma.



Gráfica 8: # hijos familia del barrio X

Hazard: Observa, la mayoría de las familias encuestadas tienen menos de 4 hijos, y son muy pocas las familias que no tienen hijos. Volviendo al tema,

aunque esa presentación de los datos es muy diciente, quiero hacer un diagrama que dé todavía más información. Pienso que podría hacer un diagrama de bloques como el que se usó en clase esta mañana, para la variable lugar de almuerzo. ¿Tú qué dices?

Chepa: Creo que el diagrama adecuado para este caso, está prácticamente insinuado por el que tú hiciste el de puntos, y en esencia es muy similar al de bloques; sin embargo, en vez de emplear rectángulos para representar cada una de las clases se puede usar segmentos de recta, cuya longitud sea proporcional a la frecuencia del correspondiente valor de la variable.

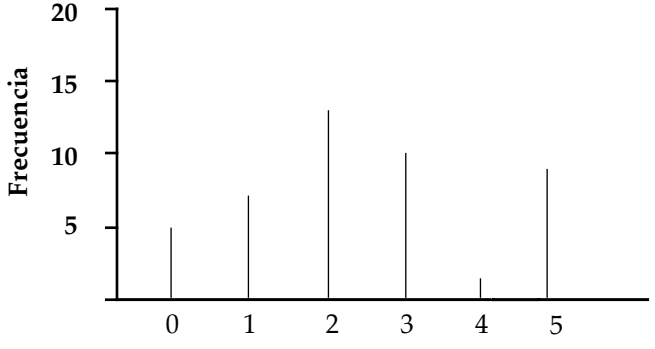
Hazard: Y, ¿cuál es la razón por la cual, para este caso, es mejor usar segmentos que bloques? ¿No es una decisión un poco caprichosa?

Chepa: Realmente puede parecer cuestión de capricho. Miremos qué diferencia hay entre las variables lugar de almuerzo y número de hijos.

Hazard: Lugar de almuerzo es una variable categórica, mientras que número de hijos es una variable cuantitativa discreta. Como sí hay una diferencia esencial entre las dos variables y convinimos usar bloques para representar las variables categóricas, entonces podemos acordar que vamos a emplear diagramas de segmentos para los casos en que trabajemos con variables cuantitativas discretas.

Chepa: Entonces, el diagrama quedaría así:

Gráfica 9



Hazard: Ahora miremos cómo está distribuida la edad de los hijos de esas 50 familias. Yo ya hice la tabla de distribución de frecuencias. Mírala. (*Hazard muestra la siguiente tabla.*)

Edad (años)	Frecuencia	Edad (años)	Frecuencia
0.4	1	7.6	2
0.7	2	8.0	2
1.5	4	10.0	5
1.7	3	10.5	2
1.9	4	10.7	4
2.0	6	11.2	2
2.3	9	11.4	3
2.6	5	11.6	1
3.0	6	13.0	2
3.5	6	13.5	1
3.7	8	14.0	2
4.2	3	14.5	1
4.6	7	15.0	1
5.0	2	15.6	2
5.3	1	16.0	3
5.7	2	16.2	2
6.0	4	16.9	1
6.2	2	17.0	2
6.8	3	17.5	1
7.0	1	17.8	2

Chepa: ¡Qué montón de valores los que asume esta variable! Me da la impresión de que para este caso no es muy adecuado hacer una gráfica como la que hicimos para representar la distribución de la variable número de hijos.

Hazard: A ver..., ¿cómo quedaría si la hiciéramos como hicimos la anterior? Si por cada valor de la variable hiciéramos un segmento de recta, tendríamos toda la información que se obtiene de la tabla. Pero, ¿qué tan diciente es esa información? Es decir, ¿qué tanto le puede significar esa gráfica a cualquier persona que pretenda describir en términos generales la situación correspondiente al problema? Realmente, ¿vale la pena dar todo el detalle de los datos? Creo que en ocasiones es preferible perder algo de información en aras de poder leer e interpretar más cómodamente, o por lo menos, más rápidamente la información, porque recuerda que, en últimas, lo que queremos es resumir la información.

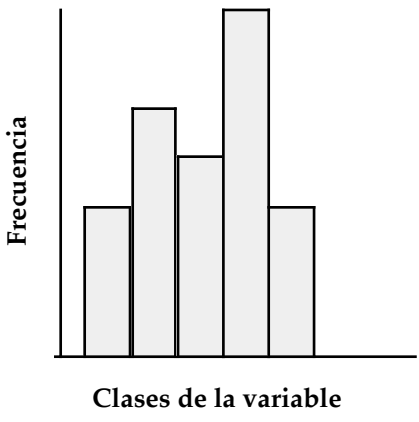
Chepa: Sí. Tal vez tú tienes razón. Además, si convinimos en emplear los diagramas de segmentos de recta para representar la distribución de una variable cuantitativa discreta, entonces debemos emplear otro tipo de gráfica en casos como el que estamos tratando, pues la variable es cuantitativa continua.

Hazard: Resumamos lo dicho hasta ahora. Para decidir cuál es una buena manera de representar gráficamente una variable cuantitativa continua debemos tener en cuenta especialmente dos condiciones: en primer lugar, debemos resumir la información y en segundo lugar, en la gráfica debe reflejarse el hecho de que la variable es continua. Por tanto, esas dos condiciones que mencioné deben marcar la pauta.

Chepa: Déjame decirlo a mí. Respecto a la primera condición que diste, hagamos grupos de edades; y con respecto a la segunda condición, en la gráfica, demos la sensación de que no hay espacios entre uno y otro grupo.

Hazard: Entonces, la gráfica debería ser de este estilo: (*Hazard hace rápidamente un dibujo y lo muestra a Chepa.*)

Gráfica 10



Chepa: Sí; parecido al diagrama de bloques, pero sin las separaciones entre las diferentes clases. Ahora, el problema que se presenta es decidir cuántas clases hay que construir.

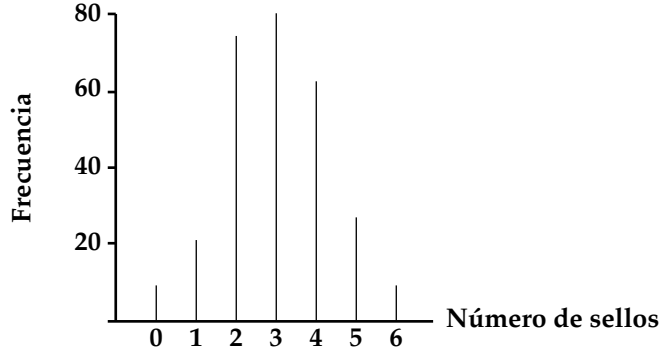
Hazard: ¡Qué lástima que no pueda quedarme para que me ayudes a resolver ese problema! Pero, creo que con lo que hemos hecho puedo salir adelante. Cuando termine de hacer el análisis de mi problema del kinder, te contaré si logré convencer o no a papá. Adiós, Chepa.

Chepa: Adiós, Hazard.

a. En el diálogo, Chepa dice "(...) Ese método de presentar la información (se refiere al diagrama de puntos, a la Gráfica 8) es mejor que dar una tabla de distribución de frecuencias; (...)". ¿Está usted de acuerdo con esa afirmación? Al justificar su respuesta mencione los pro y los contra de ese tipo de diagrama.

b. Un experimento consiste en lanzar simultáneamente seis monedas durante cierto número de veces, y cada vez anotar el número de "sellos obtenidos". El diagrama siguiente muestra los resultados. Con base en él, responda las siguientes preguntas:

Gráfica 11
Lanzamiento de 6 monedas



- ¿Cuál es el espacio muestral del experimento aleatorio definido?
- ¿Cuál fue el resultado obtenido con mayor frecuencia? ¿Cuál es el número de sellos menos frecuente?
- En el eje vertical del diagrama (el correspondiente a "Frecuencia") no está marcada la frecuencia del resultado "2 sellos" ni tampoco la de

"5 sellos". Sin embargo, observando la longitud de los correspondientes segmentos, usted puede saber dichas frecuencias. Encuéntrelas.

- ¿Qué característica importante del diagrama es la que permite responder la pregunta anterior? Explique su respuesta.
- ¿Cuál valor de la variable tiene frecuencia igual a 75?
- ¿Cuántas veces se realizó el experimento?
- ¿Cuántas veces se obtuvo menos de "4 sellos"?

c. En el diálogo, Chepa dice "(...) pues... hagamos grupos de edades; (...)" refiriéndose a una de las condiciones importantes para representar gráficamente la distribución de la variable edad. Explique por qué y para qué es conveniente hacer lo mencionado por Chepa.

d. Suponga que usted le tiene que ayudar a Hazard a establecer los grupos de edades. ¿Con qué criterios haría usted esos grupos? O, ¿cree que no es necesario fijar unos criterios para hacerlo? Explique su respuesta.

e. De acuerdo a la respuesta que dio en el item anterior, haga una tabla que muestre los grupos de edades en que usted clasifica los datos y sus correspondiente frecuencias. Además, con base en dicha tabla y teniendo como modelo el diagrama hecho al final del diálogo por Hazard, haga la representación gráfica de la distribución de las edades.

Otro resumen

En la sección "Formalicemos un poco" se mencionaron dos tipos de diagramas para representar la información contenida en una muestra: el diagrama de bloques y el circular, los cuales se emplean para distribuciones de variables categóricas.

Puesto que no todas las variables son categóricas es preciso, entonces, encontrar formas adecuadas de representar las distribuciones de esas otras variables. Y, de la identificación de esa necesidad, surge naturalmente la tarea (buscar los diagramas más adecuados para cada caso) que se realizó a través del diálogo "Otras gráficas" y de la reflexión hecha para responder las preguntas correspondientes. Ahora, vamos a hacer un resumen de las conclusiones a las que se llegó.

Se presentaron tres tipos de diagramas: el *diagrama de puntos*, el *diagrama de barras* y el *histograma*.

En caso de que la variable cuantitativa sea discreta, asuma pocos valores, el total de observaciones sea un número relativamente pequeño y el objetivo sea presentar organizadamente la información, el diagrama de puntos es una manera eficiente de representar la información contenida en una muestra. Es mejor que una tabla de distribución de frecuencias porque hace evidente la forma de la distribución de la variable. Sin embargo, quien observe este tipo de diagrama y quiera saber la frecuencia de un determinado valor de la variable deberá hacer el conteo.

> Un *diagrama de puntos* es una gráfica que se emplea para dar una idea aproximada de la forma de la distribución de una variable cuantitativa discreta. Sobre una misma recta (usualmente horizontal) se disponen en orden ascendente los posibles valores de la variable y encima de cada uno de esos valores se anotan tantos puntos como veces se repita el valor.

Aunque el diagrama de puntos se usa especialmente con variables discretas, en ciertas ocasiones puede usarse para representar la distribución de una variable continua, con el fin de tener una idea de la forma de la distribución.

El diagrama de barras es muy parecido al de bloques; tanto, que en repetidas ocasiones se emplean indistintamente. Sin embargo, nosotros haremos distinción en el uso de ellos para significar que el de bloques representa la distribución de una variable categórica, y el de barras se refiere a la distribución de una variable discreta.

> Un *diagrama de barras* es una gráfica que se emplea para representar la distribución de una variable cuantitativa discreta. Sobre un misma recta (usualmente horizontal) se disponen en orden ascendente los posibles valores de la variable y encima de cada uno de esos valores se trazan segmentos de recta cuya longitud es proporcional a la frecuencia de cada valor de la variable.

Al tener que representar gráficamente la distribución de una variable continua, la gráfica que se elija tiene que reflejar las características de dicho tipo de variable. Esas dos características se refieren a que:

- hay infinidad de valores que puede asumir la variable
- entre cualquier par de valores que asuma la variable, existe por lo menos otro valor que también puede ser asumido por la variable.

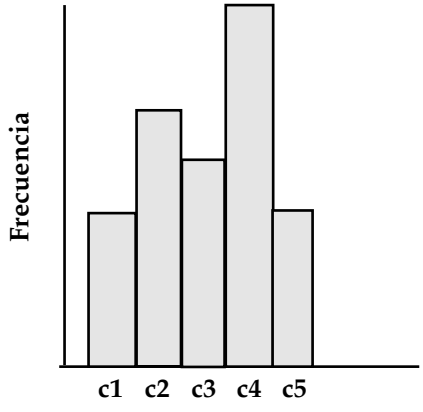
Esas dos condiciones se reflejan exactamente en la gráfica, así:

- Se usan bloques
- dichos bloques son adyacentes

La anterior descripción da lugar a gráficas como la que se muestra en la Gráfica 12. Tales gráficas reciben el nombre de *histogramas*.

> Un *histograma* es una gráfica que se emplea para representar la distribución de una variable cuantitativa continua. Está constituida por rectángulos ubicados sobre una misma recta. Cada uno de los grupos en que se clasifica la variable está representado por la base de un rectángulo; y la altura del mismo es proporcional a la frecuencia del correspondiente grupo de valores. Además tales rectángulos son adyacentes.

Gráfica 12



Como puede notar, la construcción de un histograma no presenta ninguna dificultad. El problema reside en agrupar los datos de la variable en la forma más adecuada posible. Para lograr esto es necesario tener en cuenta algunas características de los valores que asume la variable: qué tan grandes son, qué

tanta diferencia hay entre ellos, qué tantos valores hay, etc., de manera que se puedan establecer criterios de agrupación.

De lo anterior se deduce que la agrupación de los valores no es única, pues depende del criterio de quien esté haciendo la clasificación. Sin embargo, hay unas reglas prácticas que pueden servir como base para tomar la decisión.

Puesto que el objetivo de hacer un histograma es permitir un resumen de datos y facilitar la lectura e interpretación de los mismos para describir el conjunto, esas consideraciones deberán reflejarse en la gráfica.

Al tener que decidir **cuántas clases** deben hacerse, la respuesta es: ni tantas, ni tan pocas. Si se hacen muy pocas, el resumen es tal que se pierde mucha información, y si se hacen muchas clases no se está logrando un verdadero resumen. En la práctica, el número de clases suele variar entre 5 y 15.

El otro aspecto que interesa decidir es **qué tamaño** debe tener cada una de las clases. Para eso, es necesario saber qué tan grande es la variación de los datos del conjunto, y una manera eficiente de lograr esa información es calcular la diferencia entre los valores máximo y mínimo del conjunto. Por tanto, si se tiene definido el número de clases, y se quiere que todas tengan el mismo tamaño, al dividir la diferencia de los valores máximo y mínimo por el número de clases, queda determinado el tamaño que debe tener cada una de ellas. Y, con esa información sólo queda por definir cada una de las clases, determinar su frecuencia

Se quiere representar gráficamente las calificaciones de un parcial de estadística. La gráfica debe dar un cierto detalle de las diferencias entre las calificaciones de los alumnos a quienes les fue mal, regular y bien. ♣*¡O sea la distribución BRM!*♣ Dichas calificaciones son las siguientes

1.5	1.7	1.8	2.2	2.3	2.4	2.4	2.4	3.4	4.1	3.9
2.5	2.6	2.8	3.1	3.2	3.3	3.3	4.5	3.9	4.0	3.9
3.6	3.6	3.6	3.7	3.8	3.9	3.9	4.4	4.1	4.0	

Si se quiere clasificar los valores de la variable calificación, atendiendo las sugerencias dadas anteriormente, se podría llegar a algo similar a esto:

1.- Puesto que son pocos datos (32), y entre ellos no hay diferencias grandes en cuanto a su valor, hacer 6 clases puede ser suficiente.

2.- Como el mínimo valor de la variable es 1.5 y el máximo es 4.5, entonces la longitud del segmento que se quiere dividir es 4.5 - 1.5 = 3.

3.- Puesto que la longitud del segmento es 3 y se quiere hacer 6 clases, entonces el tamaño de cada clase debe ser 3 / 6 = 0.5

4.- La tabla de distribución de frecuencias agrupadas se puede presentar así:

Clase	Frecuencia
1.5 - 2.0	3
2.0 - 2.5	5
2.5 - 3.0	3
3.0 - 3.5	5
3.5 - 4.0	10
4.0 - 4.5	5

La forma como se han construido las clases presenta un problema de interpretación de las mismas. Veamos cuál es: por ejemplo, el extremo superior de la primera clase es 2.0 y es el mismo extremo inferior de la segunda clase, entonces quien interprete la tabla no necesariamente sabe en cuál de las dos clases se contó la frecuencia del valor 2.0, en caso de que ese valor sea uno de los que asume la variable. Para solucionar problemas como ese podemos convenir en que para cada clase, el extremo inferior se incluye, pero no el extremo superior. Es decir, volviendo a nuestro ejemplo, 2.0 es un valor excluido de la primera clase, pero incluido en la segunda; de esa manera se evitan posibles ambigüedades.

Con respecto a la tabla de frecuencias agrupadas del ejemplo que estamos desarrollando, si usted tuvo la precaución de totalizar las frecuencias de las diferentes clases, debió obtener 31 y no 32 como era de esperarse. ¿Qué ocurrió? Efectivamente no hay ningún error de conteo ni de suma. Sólo que en realidad, con el tamaño que se dio a cada clase quedan definidas 7 clases y no 6 como se dijo. ¿Entonces, qué hacer? **Una** forma de resolver el problema es definir la última clase como "4.0 ó más. Si eso se acepta, entonces la tabla de frecuencias queda:

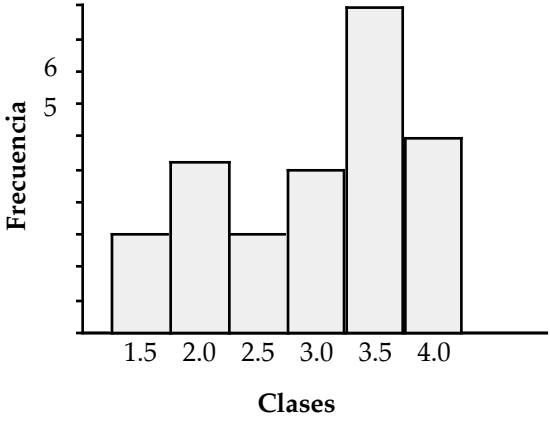
Clase	Frecuencia
1.5 - 2.0	3
2.0 - 2.5	5
2.5 - 3.0	3
3.0 - 3.5	5
3.5 - 4.0	10
4.0 ó más	6

La otra forma de solucionar el problema es adicionar una clase más, que incluya los valores de la variable que haga falta considerar.

5.- De manera que el histograma correspondiente a la tabla anterior sería el siguiente:

Gráfica 13

Distribución de calificaciones



Algunos ejercicios

1.- Hable de las diferentes formas de representar gráficamente los datos de una variable.

2.- Un grupo de politólogos está estudiando las características generales del municipio de Chía para un trabajo de investigación. Los datos[14] que se muestran a continuación corresponden al uso de los suelos en el municipio. Con base en ellos podemos ver el desarrollo del aspecto socio-económico en Chía.

Uso del suelo	Porcentaje
áreas en desarrollo	13.08
recreación y turismo	0.72
minero	4.34
agricultura	6.82
floricultura	1.41
ganadería	36.27
bosques	3.61
agro-industrial	3.82

a. ¿Cuál es la variable de estudio, de qué tipo es y cuáles valores toma?

b. Represente gráficamente la información contenida en la tabla.

3.- Un politólogo desea determinar qué tipo de mecanismo prefieren las personas para elegir a un candidato: el *tarjetón* o la *papeleta*. Uno de los asistentes del politólogo aplica una encuesta en un municipio cercano a Bogotá, Tabio, a 175 personas y encuentra que 25 no votan, 96 prefieren el tarjetón y 54 prefieren la papeleta.

14 Tomados de Pedro Gómez y Cia. grupo de consultoría.

a. ¿Cuál es la población de estudio?

b. ¿Cuál es la muestra de estudio y su tamaño?

c. ¿Cuál es la variable de estudio? Comente los resultados obtenidos por el asistente, con respecto a la variable que él pretende medir. Proponga una manera de eludir el problema que presentan los resultados.

d. Acoja la sugerencia que dio en el item anterior para elaborar una tabla de frecuencias y representar gráficamente la información.

4.- Un estudiante de Ciencia Política desea conocer cuál es el candidato para alcalde de Bogotá por el cual se inclina la mayoría de los estudiantes de la universidad donde él estudia. Para eso, toma una muestra aleatoria de 80 estudiantes de la universidad y cada uno de ellos da su voto a favor de alguno de los siguientes candidatos: el del M-19, el del Partido Conservador o el del Partido Liberal. Los resultados son los siguientes:

M-19	Lib.	Lib.	Lib.	Lib.	Lib.	Lib.	Con.	Con.
M-19	M-19	M-19	M-19	Lib.	LIb.	Con.	Con.	Con.
Lib.	Lib.	Lib.	Lib.	Con.	M-19	Con.	Con.	Lib.
M-19	Lib.	Con.	Con.	Lib.	Con.	M-19	M-19	M-19
Lib.	Lib.	Con.	Con.	M-19	M-19	M-19	Lib.	Con.
Con.	Con.	Lib.	Lib.	Lib.	Lib.	M-19	M-19	M-19
Con.	Con.	Con.	Lib.	Lib.	M-19	M-19	Lib.	M-19
Lib.	Con.	Con.	Con.	Lib.	Lib.	Lib.	N vot.	Lib.
M-19	Con.	M-19	M-19	N vot.	N vot.	M-19	M-19	

a. Diga cuál es la variable que se está midiendo y diga de qué tipo es y cuántos valores toma.

b. Realice una tabla de frecuencias que represente la información dada anteriormente.

c. Haga un diagrama que le permita ver fácilmente cuál fue el candidato con más votación.

5.- Un científico cree que el color rojo tiene un mayor grado de estimulación para las palomas, que el verde. Para comprobar su hipótesis expuso a 15 palomas a una sesión completa de condicionamiento operante que consistía en poner a las palomas en una caja que tenía dos luces (una roja y una verde) y un comedero. Las palomas debían picotear cualquiera de las luces para recibir comida. Algunas veces picoteaban la luz roja y otras, la verde. Cada sesión duraba 30 minutos y las respuestas que dieron las palomas durante ese transcurso de tiempo fueron las siguientes:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
rojo	76	49	66	82	57	67	72	84	63	59	56	64	68	81	79
verde	63	53	65	91	85	12	53	62	56	42	40	32	50	60	23

a. Identifique la población de estudio y la muestra que representa a dicha población.

b. ¿Cuál es la variable que se está midiendo? ¿De qué tipo es?

c. Haga tablas de frecuencias que representen la situación y también haga un diagrama que la represente.

6.- Un psicólogo desea establecer si los programas de televisión que presentan situaciones violentas tienen la misma influencia en niños de distinto sexo, que viven en ciudades grandes. Para ello, escogió al azar de sus 50 pacientes (25 niños y 25 niñas) 10 de cada sexo y con ellos realizó el siguiente experimento: todos los días durante un mes sometió a esos 20 niños a observar un programa violento de una hora de duración. Al finalizar el experimento, el psicólogo evaluó, por medio de un test escrito, el nivel de agresividad en cada niño. Los puntajes obtenidos en el test se dan a continuación:

	1	2	3	4	5	6	7	8	9	10
niños	88	90	76	92	86	83	67	86	90	92
niñas	93	68	65	73	89	75	62	70	64	66

a. Identifique la población de estudio y la muestra con la que se hizo el experimento.

b. Cree usted que esa muestra sea lo suficientemente representativa para estimar o predecir el comportamiento de la variable dentro de la población?

c. Haga las tablas de frecuencias correspondientes a cada grupo.

d. Para cada caso (niños, niñas) haga una gráfica que represente los resultados.

e. ¿Cree usted que el sexo sea el único factor que influye en el grado de agresividad del niño? ¿Qué otros factores pueden influir en la conducta de la cual se está hablando?

7.- El gerente de ventas de un supermercado organiza un estudio para determinar el tipo de aceite usado en la cocina. Tal estudio se lleva a cabo en la zona norte de Bogotá y se realiza con 180 familias de clase media. Los resultados fueron los siguientes: 40 familias consumen aceite de ajonjolí, 35 familias emplean aceite de soya, 15 emplean manteca de cerdo, 15 usan aceite de oliva, 50 emplean aceite de girasol y 36 familias emplean aceite de maíz.

a. ¿Cuál es la población de estudio?

b. ¿Cuál es la muestra sobre la que se va a realizar el estudio y cuál es su tamaño?

c. ¿Cuál es la variable que se está investigando? ¿De qué tipo es?

d. Construya una distribución de frecuencias.

e. ¿Cuánto da la suma de frecuencias? ¿Cómo se explica que la suma de frecuencias sea superior al número de familias del estudio?

f. Represente gráficamente la distribución de frecuencias.

8.- En una fábrica textil la producción en miles de metros de los últimos meses se presenta en la siguiente tabla. El gerente financiero está muy preocupado por la situación. ¿Cuál es la razón de ello?

enero: 3.500	marzo: 2.000	febrero: 4.000
abril: 2.500	agosto: 3.000	junio: 2.500
mayo: 3.000	julio: 2.000	

9.- La persona encargada de asignar salones a los cursos que se dictan en la universidad, quiere determinar la proporción de estudiantes que hay en cada una de las carreras que conforman la facultad de Humanidades para explicar por qué no conviene hacer la asignación de salones al azar. Después de que haya recogido la información va a presentarla en un diagrama.

a. ¿Cuál es la variable que se va a medir? ¿De qué tipo es?

b. ¿Qué tipo de diagrama es el más adecuado? ¿Por qué?

10.- Un estudiante de derecho quiere determinar la proporción de estudiantes que hay en cada una de las carreras que conforman la facultad de Humanidades para corroborar su afirmación de que los estudiantes de derecho son los que deciden en una votación pues son la mayoría. Después de que haya recogido la información va a presentarla en un diagrama.

a. ¿Cuál es la variable que se va a medir? ¿De qué tipo es?

b. ¿Qué tipo de diagrama es el más adecuado? ¿Por qué?

11.- El administrador de un supermercado está interesado en determinar si es necesario instalar más cajas registradoras en el almacén para darle una atención más rápida a la comunidad. Para el efecto, la persona encargada de hacer la investigación toma, un día cualquiera, una muestra aleatoria de 50 compradores y anota el tiempo que cada uno de ellos gasta haciendo cola para pagar sus compras. Después de que haya recogido la información va a presentarla en un diagrama.

a. ¿Cuál es la variable que se va a medir? ¿De qué tipo es?

b. ¿Qué tipo de diagrama es el más adecuado? ¿Por qué?

12.- En el curso de estadística se han realizado tres parciales. El profesor del curso cree que, en general, sus alumnos van bien pues la mayoría de ellos han aprobado los tres parciales. El profesor quiere presentar la información correspondiente en un diagrama.

a. ¿Cuál es la variable que se va a medir? ¿De qué tipo es?

b. ¿Qué tipo de diagrama es el más adecuado? ¿Por qué?

13.- Numerosos lingüistas han explicado el problema de la pluralización del verbo **haber**. Para unos, ninguna forma del verbo haber admite plural; otros explican que el uso coloquial de formas como "habían" o "han habido" se ha impuesto hasta ser aceptado; y otros argumentan que la naturaleza lingüística del verbo sí permite la forma plural.[15] A pesar de las muchas razones teóricas, la forma plural, correcta o incorrecta, sigue siendo utilizada. (La forma correcta es la singular. Por ejemplo, **había** muchos carros.)

Para profundizar en el problema, un grupo de lingüistas colombianos hizo una prueba en Cali. Realizaron entrevistas a veinte personas; diez de ellas, entre 19 y 30 años y las otras diez entre 31 y 45 años. De cada persona entrevistada se registró el empleo que ella hizo del verbo haber (según lo que interesa para este caso) en tres ocasiones tomadas aleatoriamente de la entrevista. Se codificaron con 1 las formas del verbo haber en singular y con 2 las formas del verbo haber en plural. Los resultados del corpus fueron los siguientes:

	1	2	1	2	1	1	2	2	1	1
19 - 30 años	2	2	2	1	1	2	2	2	1	1
	1	1	2	2	1	2	2	2	1	2
	1	1	1	1	1	2	2	1	1	2
31 - 45 años	2	1	1	2	1	2	1	1	2	2
	1	1	1	2	1	1	1	1	1	2

a. ¿Cuál es la población de estudio y cuál la muestra de estudio?

b. ¿Qué es un corpus?

c. ¿Cuáles variables se consideran en la investigación?

15 Bentivoglio, Paola. "Haber: ¿un verbo impersonal?" *Estudios sobre español de América y lingüística afroamericana*. Bogotá: Instituto Caro y Cuervo, 1989, pp. 61-64.

d. Para la variable que se está midiendo, elabore una tabla de frecuencias y un diagrama de bloques.

e. Considere el conjunto de datos que se le presenta según la generación. Haga un diagrama circular para cada uno de los subgrupos y compárelos.

f. ¿Cree usted que el uso del plural en el verbo haber puede ser una evolución que se da en el lenguaje a través del tiempo? (Apóyese en la respuesta dada a la pregunta anterior.)

14.- Un factor que influye notablemente en el estilo del lenguaje hablado es el ambiente en el que se encuentra el hablante. Para reafirmar la hipótesis de que existe una relación directa entre la formalidad del contexto y la formalidad del lenguaje, un profesor de lingüística de una universidad de Medellín decidió calificar el lenguaje de un grupo de alumnos suyos durante las exposiciones realizadas en clase y durante los descansos en una cafetería. Las calificaciones (que variaban entre 1 y 5, donde 1 correspondía a un lenguaje puramente coloquial y 5 correspondía a un lenguaje muy elevado o refinado), de los veinte estudiantes en los dos contextos fueron:

En la clase	3,0	3,5	4,0	4,5	2,5	2,0	3,0	4,0	4,7	3,9
	4,2	3,9	4,0	4,1	3,5	2,8	3,2	2,5	3,5	4,0
En la cafeteria	3,0	2,5	2,0	1,0	0,2	0,7	3,0	4,2	3,0	2,9
	1,0	1,5	0,9	3,0	3,4	4,0	2,7	2,8	1,5	1,9

a. ¿Cuál es el problema de estudio?

b. ¿Cuál es la muestra del estudio?

c. ¿Cuáles variables se consideran? ¿De qué tipo son?

d. Para las calificaciones obtenidas en cada uno de los dos contextos estudiados elabore una tabla de frecuencias y un histograma.

e. Compare la información ya resumida y diga si el caso particular que se está considerando parece apoyar la hipótesis planteada.

15.- El lenguaje es la herramienta propia del hombre para comunicarse con los otros miembros de una sociedad. Con el fin de comprobar el cumplimiento de la función comunicativa de su periódico en toda la sociedad cucuteña, el director de un destacado diario de la capital nortesantandereana adelantó una investigación que pretendía medir el alcance de la información contenida en el periódico con respecto al grado de comprensión que de ésta lograba el lector. Para tal efecto, se tomó una muestra de cien suscriptores del diario a quienes se pidió calificar tres tipos de escritos según el grado de dificultad de comprensión que presentaban. A continuación se especifica un poco más el estudio.

Muestra estratificada por nivel socioeconómico de los suscriptores del periódico	
Alto	20
Medio alto	30
Medio bajo	20
Bajo	30

Tipos de textos sometidos a calificación	
A	Temas de interés general
B	Temas sociales y políticos
C	Temas especializados

Escala de calificaciones para el grado de dificultad de comprensión del texto	
+2	Muy difícil
+1	Difícil
0	No muy difícil
-1	Fácil
-2	Muy fácil

Los resultados, de la calificación general a los tres textos, obtenidos según el nivel socioeconómico fueron:

Calificación según nivel socioeconómico										
Alto	0	+1	+1	0	-1	-1	-1	+2	0	0
	0	+1	+1	-1	-1	-1	0	0	+1	-1
Medio alto	-1	-2	+2	+2	0	0	+1	+1	+1	0
	-1	-2	-2	0	0	+1	+1	+2	-1	0
	-2	-1	0	0	0	+1	0	0	0	+1
Medio bajo	-1	0	+1	0	0	+1	-1	-2	+2	+1
	0	0	+1	+1	0	-1	-1	-1	-1	0
Bajo	0	0	+1	-1	-2	+2	+2	0	0	0
	+1	+1	+1	+2	0	0	0	-1	0	0
	+1	+1	+1	+2	0	0	+1	+1	0	0

a. ¿Cuál es el problema de estudio?

b. ¿Cuál es la muestra de estudio?

c. ¿Cuáles son las variables y de qué tipo son?

d. Elabore una tabla de frecuencias para las calificaciones dadas y represente esta información con un diagrama de puntos.

e. Elabore tablas de frecuencias y diagramas de barras para las calificaciones obtenidas según los cuatro niveles socioeconómicos considerados en el conjunto de datos presentado.

f. Compare los cuatro diagramas y concluya sobre la relación entre el nivel socioeconómico y el grado de comprensión.

g. ¿Puede concluir algo sobre la diferencia entre grado de comprensión y el tipo de texto presentado? Explique su respuesta.

Para terminar

En este capítulo nos hemos referido sólo a algunas de las formas de organizar y representar gráficamente la información contenida en una muestra. Eso no quiere decir que las que hemos mencionado sean las únicas, o las más útiles. Lo que ocurre es que la intención principal de este capítulo no es hacer un estudio exhaustivo de todas las formas de representación gráfica. El objetivo principal de este capítulo es mostrar cómo en las actividades de organizar y representar gráficamente los datos obtenidos, se requiere saber **qué** se quiere hacer, **para qué** y **por qué** se quiere hacer, además de conocer la naturaleza de la información, para poder encontrar, de manera natural, a partir de las respuestas que se den a esas preguntas, **cómo** se pueden hacer. En pocas palabras, el contexto, los objetivos y la naturaleza de los datos determinan la mejor forma de organizar y representar gráficamente la información.

Medidas de tendencia central

Introducción

Hasta ahora, para describir un conjunto de datos, se han empleado los diagramas. Estos son útiles para dar rápidamente una visión general del comportamiento de los valores que asume la variable. Incluso en el caso de variables categóricas, los diagramas son suficientes para dar una descripción completa. Sin embargo, para describir el comportamiento de variables cuantitativas, en general, se requiere una mayor precisión que la que puede suministrar un diagrama: es necesario que esa descripción trascienda los límites de lo visual y lo subjetivo en cuanto sea posible. Como solución a la situación planteada anteriormente, surgen las medidas numéricas. Es decir, la descripción que se quiere hacer de un conjunto de datos numéricos se puede llevar a cabo a través de unos ciertos números que dan cuenta de los aspectos importantes de la distribución de los datos del conjunto. Más exactamente, la precisión que es deseable obtener al describir el conjunto de datos numéricos se refiere a dos aspectos, cada uno de los cuales se puede traducir en una pregunta:

- ¿Existe algún valor de la variable que represente a la mayoría de los valores del conjunto de datos?
- ¿Qué tan separados están, entre sí, los diferentes valores que asume la variable?

La primera pregunta hace referencia a las llamadas *medidas de tendencia central* y la segunda, a las llamadas *medidas de dispersión*. En este capítulo nos ocuparemos de encontrar algunas de las medidas de tendencia central.

Diálogo

(En clase se da el siguiente diálogo que, de alguna manera, es continuación del diálogo que tuvieron el profesor y sus alumnos sobre "Por qué inventar herramientas".)

Profesor: Muchachos, ¿recuerdan las dos medidas que se mencionaron en clase hace algunos días, para describir un conjunto de datos?

Tahuro: Profesor, ¿está hablando del conjunto de precios de venta y del conjunto de personas que van a una reunión trimestral?

Profesor: Exactamente, Tahuro. A esos ejemplos me estoy refiriendo. ¿Cuáles fueron los valores que en cada uno de esos casos se tomaron como buenos representantes del correspondiente conjunto?

Chepa: Yo me acuerdo. Para el problema de los precios se tomó el mínimo, y para el de las personas se tomó el máximo.

Profesor: Pues bien. En la clase de hoy vamos a proponer otras situaciones con el fin de encontrar otras medidas que son muy usuales para describir conjuntos de datos numéricos. Imaginen esta situación: una persona que no conoce el reglamento de la universidad quiere tener una idea acerca del número de créditos que toman los estudiantes de ciencias sociales de segundo semestre en esta universidad. Para el efecto, recoge un conjunto de datos numéricos que se refieren al número de créditos que toman 200 alumnos de la correspondiente población. ¿Creen ustedes que el máximo o el mínimo sirvan en este caso para representar el conjunto?

Tahuro: Realmente creo que ni el máximo ni el mínimo son buenos representantes en este caso, pues corresponden a situaciones extremas en las que sólo unos pocos estudiantes habrán caído. La mayoría de los estudiantes de esa población debe estar tomando el número regular de créditos señalados en cada programa, por tanto el número que se escoja para representar al conjunto debe ser el número que aparece con más frecuencia que los demás.

Stadi Shka: Como quien dice... la moda en ese semestre es ponerse 18 créditos.

Profesor: Pues, aunque lo digas en broma, *moda* es el nombre que recibe la herramienta que mejor representa al conjunto de datos al que me estaba refiriendo. Veamos otro ejemplo: supongan que sabemos que las calificaciones de Stadi Shka, en los diez "quizes" que hemos hecho en el curso de matemáticas son: 3,8; 4,2; 2,5; 2,8; 3,8; 2,0; 4,2; 3,7; 3,5; 3,3. Si yo tuviera que elegir un número para calificar a Stadi Shka, ¿quedaría ella conforme si yo escogiera el mínimo del conjunto de sus diez calificaciones?

Stadi Shka: ¿Está loco profesor? ¿Por qué, mejor, no piensa en escoger el máximo? Sería la primera vez en mi vida que yo obtendría una buena calificación en matemáticas.

Askanio: Mejor no sueñes Stadi Shka.

Ana Liza: Ahora no vayan a comenzar a pelear. En este caso, creo que es evidente que no es buena idea escoger ni el máximo, ni el mínimo, ni la moda como representante del conjunto de datos. Debemos, entonces, encontrar una herramienta que sea eficiente para representar al conjunto de calificaciones y yo creo que es...

Profesor: No lo digas aún. Vamos a hacer el proceso para descubrir esa herramienta. Si ninguna de las herramientas que hemos inventado hasta ahora es adecuada para representar el conjunto de calificaciones, vamos a buscar una que supere los problemas o limitaciones que tienen aquéllas. Por ejemplo, la herramienta que encontremos debe tener en cuenta **todos** los datos del conjunto y no sólo algunos.

Stadi Shka: Profesor yo sé cuál es esa herramienta. Déjeme mencionarla.

Profesor: No, Stadi Shka. Yo sé que todos están pensando en la herramienta que es; pero, supongamos por un momento que no lo saben, recuerden que lo que nos interesa es el proceso que ustedes deben seguir para descubrir las herramientas que realmente no conocen. Para ello voy a guiarlos en ese descubrimiento con algunas preguntas.

Askanio: ¿Cuál es entonces la pregunta que va a formularnos?

Profesor: Si con base en las diez calificaciones de Stadi Shka que mencioné anteriormente, tuviera que dar un número que represente el rendimiento de ella en la clase de matemáticas, ¿sería adecuado dar el total de la suma de esas diez notas?

(*Por un momento, todos quedan pensativos.*)

Stadi Shka: Nunca había pensado en esa forma de calificar, pero... ahora que usted lo dice, sí, creo que sí podría calificarse de esa manera y sería más cómoda: no habría que hacer tantos cálculos aburridos.

Profesor: Sí; esa forma de obtener la calificación definitiva sería bien sencilla. Pero, es necesario determinar si esa herramienta es eficiente para asignar calificaciones que deben representar el rendimiento de una persona comparado con el de otras. En este momento yo les aseguro que la suma de todas las notas que tenga una persona en un determinado curso no siempre es una herramienta eficiente pues podría no permitir establecer comparaciones entre el rendimiento de dos o más personas y en últimas eso es lo que se busca cuando se asignan calificaciones. Ahora, ustedes deben pensar en un ejemplo que aclare lo que estoy afirmando.

Ana Liza: Supongamos que Daniel y Ricardo son alumnos del mismo profesor pero tienen la clase de matemáticas a diferentes horas. En el curso en el que está Daniel hicieron cuatro evaluaciones mientras que en el que está Ricardo hicieron cinco. Las notas de Daniel fueron 4,0; 4,2; 4,5 y 4,3; las de Ricardo fueron 3,0; 3,5; 3,8; 4,2; 4,5. Si cualquiera de nosotros debiera determinar cuál de los dos muchachos tuvo un mejor rendimiento académico, no dudaríamos en decir que...

Stadi Shka: ¡Daniel!

Profesor: ¡Ajá! La intuición nos dice que "Daniel"; si los calificáramos con la suma de sus notas, tendríamos que aceptar que fue Ricardo quien tuvo mejor desempeño. En este caso estamos seguros de tener una buena intuición y entonces tenemos que concluir que lo que creíamos que era una buena herramienta para este tipo de situaciones, en realidad no lo es.

Estrella: ¡Claro! hay algunas ocasiones en las que sí podría servir como una buena herramienta. Si las personas presentaron el mismo número de evaluaciones, forzosamente tiene mejor rendimiento aquella para la cual la suma de sus notas sea mayor.

Profesor: Aquí quiero mencionar otro de los puntos importantes. Lo que ha dicho Estrella es cierto y con seguridad ustedes lo entienden y lo aceptan, pero una herramienta es más eficiente en cuanto sea más aplicable a un gran número de situaciones análogas. De manera que cuando nosotros reinventemos una herramienta, tendremos en cuenta esta condición. Aún estamos en el problema de encontrar una buena herramienta que se pueda emplear para el propósito mencionado por Ana Liza. Concretamente, ¿cuál fue la causa por la cual no sirvió la suma de las notas como una herramienta suficientemente "universal"?

Estrella: Que el número total de calificaciones no es el mismo.

Profesor: ¡Perfecto! Por tanto, el problema de la particularidad de la herramienta debería quedar resuelto si la herramienta misma tiene en cuenta ese factor.

Estrella: Sí, yo creo que no debe importar cuántas evaluaciones haga, sino el desempeño académico del estudiante.

Askanio: Muy bien, pero ¿cómo se determina el desempeño académico al que tu te refieres?

Estrella: Veámoslo con un ejemplo. Si en tres evaluaciones, las calificaciones de Askanio son, por ejemplo, 4,0; 4,0; y 4,0 es razonable calificarlo globalmente con 4,0. Si las tres calificaciones de Tahuro son 2,5; 4,5; y 5,0 entonces se tendría que:

(*Estrella escribe en el tablero.*)

notas Askanio	4,0	4,0	4,0
notas Tahuro	2,5	4,5	5,0
diferencia	1,5	-0,5	-1,0

por un lado, en la primera evaluación Askanio le lleva a Tahuro una ventaja de 1,5; en la segunda, es Tahuro quien tiene una ventaja sobre Askanio de 0,5; y en la tercera, también es Tahuro quien tiene una ventaja de 1,0 sobre Askanio; de modo que al final de cuentas ninguno aventaja al otro, por tanto, si la calificación de Askanio es 4,0 también lo ha de ser la de Tahuro.

Chepa: Ya veo; el desempeño académico que se refleja en la calificación global se basa en hacer de cuenta que el alumno no tuvo altibajos en sus calificaciones, o sea, en suponer que en todas las evaluaciones obtuvo la misma calificación.

Stadi Shka: ¡Claro! Y eso se logra repartiendo la suma de todas las calificaciones entre el número de calificaciones obtenidas.

Profesor: Pues bien. Hemos descubierto la herramienta que estábamos buscando; es el promedio o lo que, normalmente, llamaremos la *media*. Aunque el concepto es intuitivo y sencillo, pues sólo se requiere sumar los datos y dividir por el número de datos, me parece interesante que utilicemos este ejemplo

de herramienta para encontrar una *fórmula* que la represente y de esa manera ustedes se den cuenta de que la fórmula *no es más que una forma reducida de escribir —o describir— la herramienta.* ¿A alguien se le ocurre una idea?

Chepa: A mí se me ocurre, profesor. Por ejemplo, tomemos el caso de...

Profesor: No Chepa. Si lo que estamos buscando es la fórmula para la herramienta, no nos sirve describirla con un ejemplo. Necesitamos una forma de describirla que sea aplicable a todos los casos posibles *y no sólo* a un ejemplo particular.

Stadi Shka: A mí me da la impresión de que el problema es encontrar una *forma general* de identificar los elementos que intervienen en la herramienta. Y en este caso, afortunadamente, no hay sino dos elementos: los datos y el número de datos.

Profesor: Exacto. ¿Se les ocurre alguna manera *general* de representarlos?

Tahuro: Para el número de datos, el problema me parece sencillo. Como es un número, lo podemos representar por una letra. Por ejemplo, n. Se tiene entonces que n representa el número de datos y como estamos en el caso general, n representa cualquier número entero no negativo.

Chepa: ¡Ah! Ahora sí entiendo qué quería decir el profesor cuando me pidió que no explicara la cuestión por medio de un ejemplo. Pero, aun así me queda un poco complicado imaginarme cómo podemos representar de *manera general* el conjunto de datos. Para comenzar, no sabemos cuántos son.

Stadi Shka: Claro que sabemos cuántos son, Chepa: son n. Creo que tenemos que representarlos también por medio de letras. Sin embargo, veo que tenemos un problema y tal vez a eso te referías tú, Chepa. No podemos representarlos por letras cualesquiera como **A, B, C,** etc. porque como no sabemos *exactamente* cuántos son, no sabemos cuándo tendríamos que acabar de hacer la lista.

Tahuro: Pero para eso hay una solución. Utilicemos una misma letra y a cada dato lo diferenciamos con una marca, con un número. De esa forma, podríamos hablar del *primer dato,* del *segundo dato,* del *tercer dato,* y así sucesivamente hasta el *enésimo dato* que sabemos que es el último.

Chepa: Este Tahuro, cuando le da, le da. ¡Qué genio! Ahora sí entiendo. Pode-

mos representar los datos, sencillamente, por x_1, el primero; x_2, el segundo; x_3, el tercero, y, así sucesivamente, hasta llegar al enésimo que lo representaríamos por x_n.

Profesor: Muy bien Chepa. ¿Ves que la cosa no es tan difícil? Pero, ustedes no han respondido la pregunta que les hice. ¿Cómo podemos representar la *media* de un conjunto de datos con una fórmula?

Estrella: Profesor, a mí se me ocurre una especie de fórmula, pero como no tiene símbolos matemáticos, no sé si realmente sea una fórmula.

Profesor: A ver Estrella, ¿qué se te ocurre?

Estrella: Pues a mí se me ocurre que uno podría decir que:

media = suma de todos los datos dividida por el número de datos.

O lo que es lo mismo, para resumir la cosa:

$$\text{media} = \frac{\text{suma de todos los datos}}{\text{número de datos}}$$

Tahuro: Sin embargo, fíjate Estrella que ya tenemos unas maneras de *resumir* lo que estás diciendo. Por ejemplo, sería lo mismo decir:

$$\text{media} = \frac{\text{suma de todos los datos}}{n}$$

Stadi Shka: De acuerdo Tahuro. Pero podemos resumir aún más. Podríamos escribir:

$$\text{media} = \frac{x_1 + x_2 \ldots + x_n}{n}$$

Chepa: ¡Un momento! Me acabo de acordar de un símbolo que vimos al principio del curso de matemáticas. Y creo que sirve aquí como medio para resumir aún más la cosa. Miren, yo creo que la fórmula es:

$$\text{media} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Profesor: Muy bien, muchachos. Ahora consideren la siguiente situación. Carlos, un estudiante de la Universidad, entra a una librería para averiguar los precios de siete libros que debe comprar. Los precios son $7.200, $6.500, $7.300, $6.000, $6.000, $8.000 y $18.900. Cuando llega a su casa le dice a su papá que en promedio cada libro vale $8.557. ¿Ustedes qué opinan de la afirmación de Carlos?

Stadi Shka: Un momento, profesor. Ya casi tengo el promedio. Sí, Carlos tiene razón.

Askanio: Profesor, repítame los precios, ¿quiere?

Profesor: Mira, son: $6.000, $6.000, $6.500, $7.200, $7.300, $8.000 y $18.900.

Askanio: Me parece raro tener que aceptar que el precio promedio de cada libro sea $8.557. Pero, en fin, si ustedes lo aseguran.

Ana Liza: Un momento, Askanio. No cerremos este asunto tan rápido. Yo estoy de acuerdo contigo: a pesar de que el valor que calculó Stadi Shka es correcto, creo que en este caso no tiene sentido emplear el promedio para representar a todos los elementos del conjunto. En realidad, $8.557 no representa a ninguno de los diferentes valores del conjunto de los siete precios. Por tanto, opino que la afirmación que hizo Carlos es cierta, pero carece de sentido en el caso dado.

Profesor: Lo que han dicho Askanio y Ana Liza es muy cierto. Observen qué características tienen los valores de ese conjunto de precios y traten de determinar cuáles son las razones por las cuales no es adecuado en ese caso, hablar de promedio.

Tahuro: Puede ser porque hay pocos datos, sólo hay dos iguales y la diferencia entre algunos de ellos es grande; en particular la diferencia entre el mayor y el menor es enorme. Además, el valor máximo es un valor "raro" en el conjunto.

Chepa: Bueno. Y, en casos como este en el que ni la moda ni la media aritmética son valores adecuados para representar el conjunto, ¿se puede encontrar alguna otra medida que permita dar alguna idea de lo que ocurre en el centro de la distribución?

Profesor: En casos como ese, podríamos recurrir entonces, a determinar cuál es el dato central del conjunto; es decir, después de haber ordenado los datos

de menor a mayor, podríamos escoger aquel valor del conjunto antes del cual y después del cual hay igual cantidad de datos.

Askanio: En el caso que estamos analizando, ese número es $7.200. ¿Verdad?

Ana Liza: Sí, y como medida que está en el centro del conjunto expresa que en dicho conjunto hay tantos valores menores que 7.200 como valores mayores que 7.200.

Askanio: Y, aunque ese valor no represente muy bien al conjunto de los precios en el que estamos interesados, de todas maneras, sí es más representativo del conjunto que la moda y que la media aritmética.

Stadi Shka: Y, ¿qué nombre recibe esa herramienta?

Profesor: La *mediana* del conjunto. Hasta aquí dejamos por hoy. Adiós muchachos.

a. Explique qué entiende por medidas de tendencia central.

b. Nombre las medidas de tendencia central presentadas en el diálogo, defínalas y dé un ejemplo en el que sea pertinente emplear cada una de ellas.

c. Si la variable es categórica, ¿tiene sentido calcular alguna de las medidas de tendencia central? ¿Cuál? Explique su respuesta.

d. Las calificaciones de Estrella en el curso de estadística están dadas en el siguiente conjunto: {3,5, 4,5, 3,2, 4,5, 3,8, 4,2, 4,2, 4,2}. Para dicho conjunto, calcule la media aritmética, la mediana y la moda. ¿Cuál de esas tres medidas es la más representativa en este caso? Explique su respuesta.

e. Si tiene que calcular la mediana de un conjunto que tiene un número par de elementos, ¿cómo hace el cálculo? Dé un ejemplo.

f. Considere el conjunto {1, 2, 3, 4, 5, 6, 7, 8, 9} como la población de datos de un cierto estudio. De esa población, obtenga diez muestras **aleatorias** de tamaño 5 (suponga que puede construir las muestras con sustitución y con orden). ♣*Según eso, {1, 1, 1, 2, 3} es una de tales muestras y además esa muestra*

es diferente de {1, 3, 2, 1, 1}.♣ Para cada una de esas diez muestras obtenga la media aritmética, la mediana y la moda. Con base en esas repuestas diga cuál de las tres medidas de tendencia central tiene mayor variación de muestra a muestra.

g. Suponga que la distribución de una variable es tal que su media aritmética, su mediana y su moda son el mismo valor. ¿Qué característica debe tener la gráfica de esa distribución? Para dar su respuesta recurra a ejemplos en los cuales se cumplan la condiciones que se están imponiendo. Una vez que los tenga, haga las correspondientes gráficas y con base en ellas dé su respuesta.

h. Explique qué significa para usted la siguiente afirmación:

> "La media aritmética es una medida muy sensible a los valores muy grandes o muy pequeños."

i. Si para dos muestras de datos sacadas de la misma población de datos, usted conoce las correspondientes medias aritméticas, ¿puede a partir de ellas calcular la media aritmética del grupo que reúne a las dos muestras? ¿Cómo?

j. Responda la pregunta anterior para el caso de la mediana y para el caso de la moda.

Un resumen

El problema central que motiva el desarrollo de este capítulo es la necesidad de describir, objetivamente y de manera más o menos sencilla, un conjunto de datos numéricos. Para hacer esa descripción hemos recurrido a definir unas ciertas medidas que contribuyen a dar la imagen de la correspondiente distribución porque determinan valores especiales del conjunto. Tales medidas se llaman *medidas de posición*. Y, ese nombre es muy significativo: esas medidas señalan valores de la distribución que se destacan de los demás por el "lugar" que ocupan dentro de ella.

Es así como dos de esas medidas —que no fueron mencionadas en el diálogo de este capítulo por estar ya presentadas en el capítulo titulado "¿Qué vamos a hacer y cómo lo vamos a hacer?"— son el mínimo y el máximo del conjunto. Aunque esas dos medidas son muy sencillas y no aportan mucho a la descripción del conjunto de datos, de todas maneras son útiles por cuanto señalan cuáles son los valores extremos entre los cuales se encuentran todos los posibles valores de la variable.

Además de esas dos medidas de posición, están las medidas de tendencia central. Son valores de la variable alrededor de los cuales se agrupan gran cantidad de valores de la variable; son valores especiales por diferentes razones. Por ejemplo, la moda es especial porque es el valor de la variable que aparece con más frecuencia en la distribución. La mediana es especial porque es el valor de la variable que está en el centro del conjunto cuando los datos se han dispuesto de menor a mayor.

Más formalmente se pueden establecer las siguientes definiciones:

La *moda* de un conjunto de datos es el valor observado con mayor frecuencia.
La *mediana* de un conjunto de *n* datos es el valor que está ubicado en el centro, cuando se han ordenado los datos de menor a mayor o viceversa y si *n* es un número impar. Si n es par, la mediana es el promedio de los dos valores ubicados en el centro de la distribución.
La *media aritmética* o simplemente la *media* de un conjunto de datos es la suma de los datos dividida por la cantidad de datos del conjunto. Es decir, media $= \frac{1}{n} \sum_{i=1}^{n} x_i$

En este punto del desarrollo del texto conviene definir dos conceptos fundamentales: *parámetro* y *estimador de parámetro*. Los parámetros son valores que se refieren a poblaciones, en tanto que los estimadores de parámetros son valores que se refieren a las muestras. Así, por ejemplo, si P = {1, 2, 3, 4, 5} es una población y M = {2, 2, 4, 5} es una muestra extraída de P, con sustitución, se tiene que:

$(1 + 2 + 3 + 4 + 5)/5 = 3$, —la media de la población P— es un parámetro

$(2 + 2 + 4 + 5)/4 = 3{,}25$, —la media de la muestra M— es un estimador de la media de la población

Aunque la esencia del concepto es la misma en el caso de la media de la población que en el caso de la media de una muestra, en estadística se requiere distinguir los dos valores y ese hecho se refleja en la notación empleada. En términos generales, los parámetros se denotan con letras griegas y los estimadores de parámetros con letras de nuestro alfabeto. Particularmente, la media de la población se designa con el símbolo μ y la media muestral se designa con el símbolo $\bar{x}$.

Algunas consideraciones generales sobre la media, la mediana y la moda

La moda es la única medida que sirve tanto para el caso de variables categóricas como para variables cuantitativas, puesto que su definición no exige ni ordenar los valores de la variable, ni hacer operaciones matemáticas con ellos. Por ejemplo, si se está trabajando la distribución de la variable nacionalidad para un determinado conjunto de personas, no tiene ningún sentido hablar del promedio o de la mediana de esa distribución; en cambio sí tiene sentido hablar de la moda.

Si se trabaja la distribución de una variable cuantitativa, en principio, tiene sentido calcular la media, la moda, y la mediana; pero, para efectos prácticos, puede no tenerlo. Entonces, hace falta desarrollar un cierto criterio para decidir en casos particulares, cuál es la mejor medida de tendencia central. Ese criterio está relacionado con por lo menos dos factores: en primer lugar, tiene que ver con los objetivos que hay detrás del estudio que se está realizando. En segundo lugar, tiene que ver con los datos mismos: qué tan homogéneos son, qué tan típicos son, etc. Con respecto a lo expuesto anteriormente, con frecuencia, si la moda y la media de una distribución difieren mucho, es preferible usar la moda que la media.

La media es muy sensible a valores extremos muy grandes o muy pequeños. Esto quiere decir que si en la distribución hay valores evidentemente atípicos, la media no necesariamente es un buen representante de todos los datos. En cambio, la mediana no es sensible a valores extremos muy grandes o muy pequeños. En realidad, su definición es independiente del valor mismo de los datos. La moda tampoco es sensible a valores extremos muy grandes o muy pequeños, a menos de que tales cambios afecten su propio valor.

Ejercicios

1.- Se sabe que en un municipio de la Sabana de Bogotá, Cota, hay un potencial electoral de 13.875. Unos estudiantes de Ciencia Política desean saber cuál es el promedio de edad de los votantes en dicho municipio, cuál de las edades es la que presenta mayor afluencia a las urnas el día de las elecciones. Esto con el propósito de determinar cuál es la edad de los votantes que con mayor frecuencia elige a sus representantes.

Los estudiantes obtuvieron los siguientes datos: de 9.680 personas que votaron las pasadas elecciones presidenciales (mayo de 1990) se tomó una muestra de 100 personas, teniendo en cuenta su edad dentro del proceso electoral. El registro de las edades de esas personas se da a continuación:

18	18	23	20	19	70	19	47	32	43	43	45	19	20	24
19	21	21	26	20	65	19	71	21	33	18	29	40	19	52
24	23	44	20	34	24	47	29	39	40	21	55	30	21	18
45	41	61	35	35	23	27	33	21	18	19	34	61	37	18
63	38	35	46	41	34	23	36	19	20	26	40	28	20	34
29	39	28	50	48	20	23	20	37	24	32	30	19	30	27
29	53	27	44	32	21	43	23	21	37					

a. Identifique la población de estudio y la muestra de estudio.

b. ¿Cuál es la variable que se quiere medir? ¿De qué tipo es?

c. Elabore una tabla de frecuencias para organizar la información y con base en ella haga un diagrama que le permita mostrar claramente cuál es la edad con mayor nivel de votación.

d. ¿Cuál es la medida de tendencia central más apropiada para el estudio de los politólogos? Explique su respuesta.

2.- Se realizó una investigación para conocer por cuál de las listas de la pasada elección a Asamblea Nacional se inclinaron más los estudiantes de la Universidad X de Bogotá. Se tomó una muestra de 50 estudiantes de diferentes

carreras y cada uno de ellos dio el número de la lista por la cual votó. Había 118 listas inscritas. Se encontraron los siguientes resultados:

09	24	02	09	09	27	09	113	02	09
113	113	24	09	24	99	09	24	90	09
09	24	90	38	113	09	113	09	90	24
24	36	90	36	24	13	38	24	09	90
90	27	32	73	90	32	90	90	24	09

a. Identifique la población del estudio y la muestra de estudio.

b. ¿Cuál es la variable que se está midiendo y de qué tipo es?

c. Haga una tabla de frecuencias que represente la información dada. Elabore una gráfica que muestre claramente cuál es la lista con mayor votación.

d. De las medidas de tendencia central, ¿cuál considera usted que sea la más apropiada para observar por qué lista se inclinó la mayoría de los estudiantes? Explique por qué y calcúlela.

e. ¿Tiene sentido calcular las otras medidas de tendencia central? ¿Por qué?

3.- En un experimento con la caja de Skinner, una rata debe oprimir una palanca un determinado número de veces para recibir comida. Se sometió a una rata a diez sesiones en la caja de Skinner y se obtuvieron los siguientes resultados:

Sesión	1	2	3	4	5	6	7	8	9	10
# respuestas	53	67	52	76	58	63	49	62	84	75

a. Identifique la población de estudio y la muestra.

b. ¿Cuál es la variable que se está midiendo?

c. ¿Qué objetivos podría tener el estudiante al realizar el experimento?

d. Haga una tabla de frecuencias después de haber agrupado los datos en cinco grupos: 40-49, 50-59, etc.

e. De acuerdo a la tabla anterior haga un histograma.

f. ¿Qué medida de tendencia central podría ser la mejor para describir los resultados obtenidos?

4.- Ana Liza investiga el precio de cierto artículo en veinte almacenes de Bogotá y encuentra la siguiente información: (en pesos)

82	60	60	74	87	74	82	74	82	82
60	68	74	74	87	68	74	68	68	82

a. ¿Cuál es la variable que está investigando Ana Liza? ¿De qué tipo es?

b. Presente una tabla de distribución de frecuencias.

c. Haga un diagrama de puntos para representar la distribución.

d. Calcule tres medidas de tendencia central que describan la distribución. ¿Cuál de ellas es más representativa en este caso?

e. Agrupe los datos en seis clases (todas del mismo tamaño) y represente gráficamente esa información.

5.- De 400 estudiantes cuya estatura media es 1,57 metros, 250 son mujeres. Si la estatura media de las mujeres es 1,54 metros, ¿cuál es la estatura media de los hombres?

6.- Enuncie un problema para el cual la media aritmética de los datos sea cero.

7.- Las calificaciones de los alumnos de un curso de estadística en el examen final se presentan en la siguiente tabla de frecuencias agrupadas:

Clase	Frecuencia
[1,5, 2,0)	3
[2,0, 2,5)	6
[2,5, 3,0)	8
[3,0, 3,5)	15
[3,5, 4,0)	12
[4,0, 4,5)	7
[4,5, 5,0]	4

a. Haga un diagrama que represente la distribución de la variable.

b. ¿Cuántas clases se definieron? ¿Cuántos alumnos presentaron el examen?

c. Invéntese una manera de calcular la calificación promedio. (Tenga en cuenta que hay siete clases pero **no** siete alumnos: hay muchos más. Además, por ejemplo, sólo tres de ellos obtuvieron nota inferior a 2,0, mientras que doce de ellos sacaron calificación no inferior a 3,5 e inferior a 4,0). ♣ *Observe que usted sabe, por ejemplo, que hay ocho alumnos cuya calificación está en el intervalo [2,5, 3,0), pero no sabe exactamente cuál fue la calificación de cada uno de ellos. Por tanto, tiene que elegir una calificación que represente esas ocho calificaciones.*♣

d. Invéntese una manera de calcular la calificación mediana y la moda. Encuentre esas medidas.

8.- Las diferencias en la riqueza del lenguaje de niños en edad escolar pueden ser estudiadas desde el punto de vista sociolingüístico, es decir, introduciendo factores sociales en la explicación de fenómenos lingüísticos. Un estudio realizado en una escuela pública de Baltimore (EUA), sobre el porcentaje de verbos utilizados por veinte niños blancos y veinte niños negros en conversaciones de 30 minutos, arrojó los siguientes resultados:[16]

Niños blancos		**Niños negros**	
% de verbos	frecuencia	% de verbos	frecuencia
19,4	1	12,0	1
20,0	2	13,9	2
22,5	1	14,3	2
23,2	2	15,0	4
26,7	2	18,6	2
28,1	5	20,6	2
28,7	1	23,7	2

16 Datos tomados de Entwisle, Doris. "Developmental Sociolingüistics: Inner-city Children". *Advances in the Sociology of Language,* Joshua Fishman ed. Paris: Mouton Publishers, 1972, p.438. Los datos corresponden al porcentaje de verbos empleados por el niño, sobre el total de palabras de su conversación.

Niños blancos		Niños negros	
% de verbos	frecuencia	% de verbos	frecuencia
29,8	2	24,5	2
31,8	3	29,0	2
34,4	1	30,0	1

a. ¿Cuáles variables considera la investigación? ¿De qué tipo son?

b. Compare el comportamiento de la variable en cada grupo por medio de histogramas. (Recuerde que la comparación sólo tiene sentido si hay uniformidad de escalas, número de clases y tipo de frecuencia en los histogramas que se van a comparar.)

c. Los investigadores afirman que los niños blancos de la muestra tienden a usar frecuentemente mayor cantidad de verbos al expresarse que los niños negros. Compruebe esta hipótesis, ayudándose con la medida de tendencia central que usted crea más relevante para la hipótesis. Justifique su respuesta.

d. El rector de la escuela estudiada afirmó que no había diferencia entre el uso promedio de verbos por niños blancos y por niños negros de la muestra. ¿Es verdadera la afirmación del rector? Justifique.

e. ¿Qué factores (sociales, de la investigación misma, etc.) influyen en las posibles diferencias entre los dos grupos de estudio?

f. ¿Qué validez tiene comparar estos dos grupos de niños? Piense en sus costumbres, su lenguaje, su status social, etc.

9.- La influencia de los medios de comunicación sobre el hombre moderno es un elemento importante en el análisis de los cambios y variaciones que presenta una lengua en un contexto social determinado. Con el fin de estudiar el influjo de la televisión en la adopción de expresiones o formas lingüísticas incorrectas, se realizó una investigación que comparaba el lenguaje usado en los cinco programas de T.V. de mayor rating entre jóvenes de 15 a 20 años con el lenguaje de 60 jóvenes pertenecientes a este grupo de edad. Para esto, se realizaron entrevistas de 1 hora y se contaron las palabras y expresiones propias de los personajes de T.V. usadas por los jóvenes en cuestión. Los resultados arrojados por los *corpora* son:

10	15	8	2	9	10	15	22	25	30
22	20	21	25	28	30	31	30	29	17
20	21	35	40	42	45	33	25	22	20
20	21	22	23	41	38	20	19	20	18
27	29	30	33	27	20	19	20	34	20
20	31	16	8	40	43	41	20	25	23

a. ¿Cuáles son los *corpora* de la investigación?

b. La investigación planteaba como una de sus hipótesis que existe una influencia de la T.V. sobre el lenguaje si el promedio de palabras o expresiones iguales era más de 30. Con los datos disponibles verifique la hipótesis.

c. Se esperaba que de las 50 palabras y expresiones encontradas en el corpus obtenido de los cinco programas de T.V., al menos veinte fueran usadas por el 35% de las personas entrevistadas para que se pudiera pensar que existe influencia de la T.V. en el lenguaje. ¿Los resultados obtenidos confirman la hipótesis?

Medidas de dispersión

Introducción

En el capítulo anterior se definieron tres medidas que permiten indicar la ubicación del centro de una distribución y que, por tanto, contribuyen a la descripción del correspondiente conjunto de datos. Pero, para lograr una imagen completa de cómo es la distribución que se está describiendo, es necesario saber cómo se distribuyen los datos. Retomamos, entonces, la segunda pregunta formulada al inicio del capítulo anterior:

- ¿Qué tan separados están, entre sí, los diferentes valores que asume la variable?

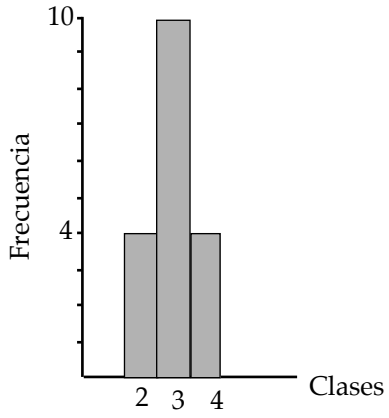
El objetivo principal de las actividades que se hagan en este capítulo será encontrar una forma adecuada de responder a tal pregunta, es decir, encontrar una buena manera de medir esa dispersión.

El rango

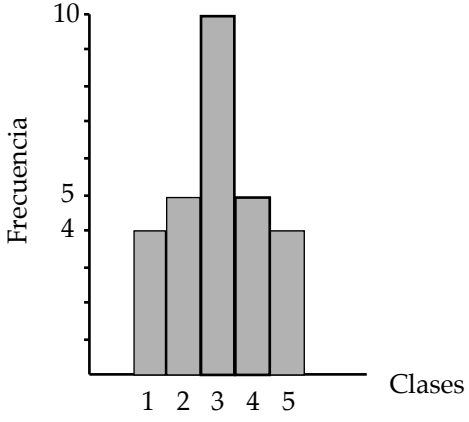
Considere la siguiente situación. En dos secciones del curso de estadística se aplicó el mismo examen, pues se quiere comparar el desempeño de las dos secciones. Las calificaciones obtenidas por los dos grupos se muestran en los siguientes histogramas.

a. Observe los diagramas correspondientes a la distribución de frecuencias de las calificaciones de cada sección y, con base en esa observación, determine si alguna de las secciones tuvo mejor desempeño que la otra. ♣*Le doy una ayuda: yo preferiría ser uno de los que presentó el examen en la sección 1.*♣ Explique su respuesta.

Gráfica 1
Desempeño de la sección 1



Gráfica 2
Desempeño de la sección 2



b. ¿Sirve, para comparar las dos distribuciones, emplear la media de cada una de ellas? ¿Por qué? ¿Qué significa, en términos de desempeño, que la media de los dos conjuntos de datos sean iguales?

c. ¿Por qué, a pesar de que la media de las dos distribuciones es la misma, ellas no reflejan que las dos secciones del curso hayan tenido el mismo

rendimiento en el examen? ¿En cuál de las dos secciones, los estudiantes tuvieron un desempeño más uniforme?

Si usted contestó que los estudiantes de la sección 1 tienen un desempeño más uniforme que los de la sección 2, eso significa que usted ha notado que lo que diferencia a los dos diagramas es la *dispersión* de los datos. Dicho de otra manera, las calificaciones de la sección 2 presentan más variabilidad que las de la sección 1. El diagrama que representa la situación de la sección 2 muestra que las calificaciones son más dispersas; es decir, que el desempeño del curso no fue tan homogéneo, como en el caso de la sección 1, para el que el diagrama resulta más compacto. Ahora intentemos descubrir una herramienta para medir esa dispersión.

d. Halle las notas máxima y mínima obtenidas en cada una de las dos secciones. Utilice el valor máximo y el mínimo de una distribución para inventar una herramienta que le permita argumentar por qué la dispersión de los datos de la sección 2 es mayor que la dispersión de los datos de la sección 1.

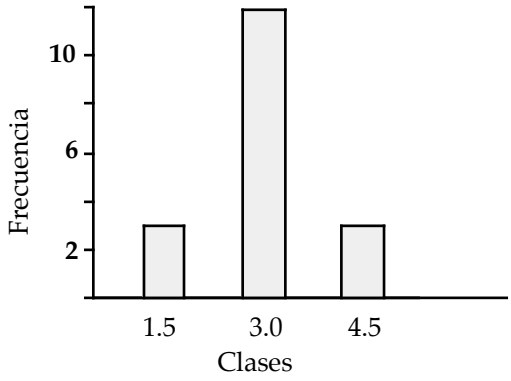
Note que los datos de la sección 2 se encuentran más dispersos, puesto que la nota mínima obtenida en la sección 2 es menor que la nota mínima de la sección 1, y además, la nota máxima obtenida en la sección 2 es mayor que la nota máxima de la sección 1. En otras palabras, las notas de la sección 2 varían en un *intervalo más grande de valores*. En este caso es muy fácil encontrar una medida que permita medir y comparar la dispersión de los datos de los conjuntos. Tal medida se llama *rango* y se define así:

> El tamaño del intervalo en el cual varían los elementos de un conjunto de datos numéricos es lo que se conoce con el nombre de *rango* y se define como la diferencia entre el mayor y el menor valor de dicho conjunto.

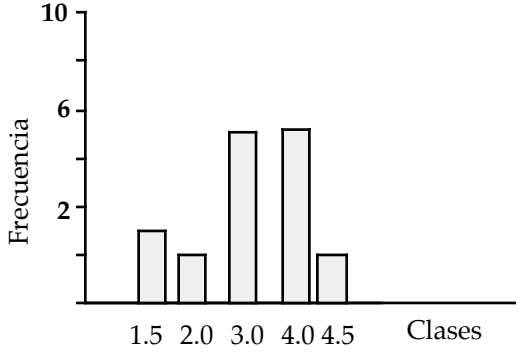
e. ¿Cree usted que el rango le sirve para comparar eficientemente la dispersión entre los elementos de cualquier par de conjuntos? Busque un caso en el que el rango no le sirva para medir y comparar la dispersión de dos conjuntos de datos.

Se ha encontrado una herramienta fácil de calcular para medir la dispersión de los datos. Pero, ¿es universal? Es decir, sirve para medir la dispersión en todos los casos? Veamos el siguiente ejemplo, usando datos de calificaciones de un examen:

Gráfica 3
Calificaciones de la sección 1



Gráfica 4
Calificaciones de la sección 2



f. ¿Cuál gráfica, refleja mayor dispersión de los datos? En otras palabras, ¿en cuál curso cree usted que fue menos homogéneo el desempeño de los estudiantes? Justifique su respuesta.

g. Calcule el rango para cada uno de los conjuntos de datos. ¿Sirve, en este caso, el rango para comparar adecuadamente la dispersión? ¿Por qué?

Se ha visto que, en este caso, el rango es el mismo para las dos distribuciones de calificaciones. Por otro lado, tenemos que aceptar que la dispersión de los dos conjuntos no es la misma. ♣*La distribución de las calificaciones de la sección 1 es menos dispersa que la de las calificaciones de la sección 2.*♣ Por tanto, en este caso el rango es un

número que **no** refleja las diferencias que, según la intuición y la observación, son evidentes entre las dos distribuciones. Y, esa situación nos exige seguir buscando una medida de dispersión que supere las limitaciones que tiene la que hemos encontrado.

h. Retome la definición de rango y determine cuál es la causa por la cual ésta no es una buena medida de la dispersión de todos los datos de la distribución.

Se ha encontrado que un problema que presenta la definición del rango como medida de la dispersión de un conjunto de datos es que depende exclusivamente de dos de los datos del conjunto y en cambio, la dispersión depende de todos los datos de la distribución. Descubramos, pues, una herramienta que tenga en cuenta *todos* los datos.

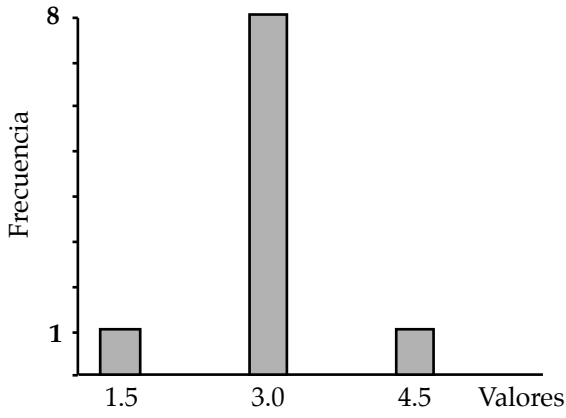
La varianza y la desviación estándar

Es claro que se necesita inventar una herramienta más fina, que remedie la deficiencia del rango; es decir, una herramienta que tenga en cuenta todos los valores del conjunto y no sólo los valores extremos.

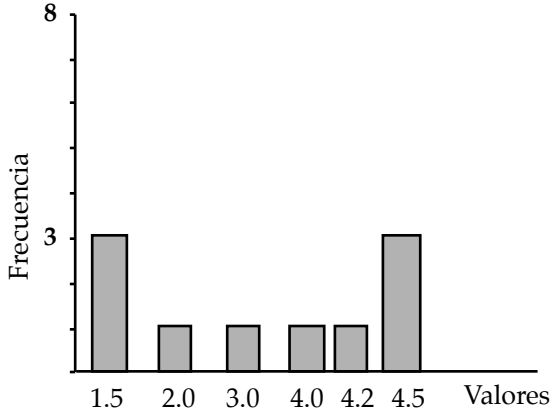
Pero, ¿cómo medir la dispersión con una herramienta que verdaderamente represente la sensación de dispersión? Note que la dispersión de los datos se refleja en que las gráficas resultan concentradas en pocos valores. A menor dispersión, los datos se encuentran concentrados en pocos valores, como se observa en los diagramas de barras de la siguiente página.

a. Suponga que, para cada una de las situaciones representadas en las gráficas siguientes, se va a realizar lo siguiente: primero, **calcular, para cada valor, la distancia que hay de él a un dato fijo**; segundo, **sumar esas distancias**; y, tercero, **emplear dicho resultado como medida de dispersión de la correspondiente distribución**. ¿Qué relación hay entre esas sumas de distancias? Es decir, ¿en qué caso resultará mayor esa suma de distancias?

Gráfica 5



Gráfica 6



b. Generalizando la respuesta a la pregunta anterior, ¿qué relación hay entre la suma de distancias a un dato fijo para una distribución muy dispersa y la correspondiente suma de distancias para una distribución menos dispersa?

En este momento debe ser claro, que para medir la dispersión de un conjunto de datos, es necesario tener en cuenta la distancia que hay de cada uno de los datos del conjunto a otro dato, que se va a tener como referencia. El problema consiste en determinar cuál es la mejor referencia. Veamos si conviene que el mínimo sea tal referencia.

c. Considere cada una de las siguientes distribuciones:

Distribución 1: 1, 2, 3, 4

Distribución 2: 5, 6, 7, 8

¿Alguna de las dos distribuciones es más dispersa que la otra? ¿Por qué?

d. Halle el mínimo de cada una de las dos distribuciones y calcule la distancia que hay de cada uno de los datos de la primera distribución al mínimo de dicha distribución (diferencia entre un valor de la distribución y el mínimo de la misma) y haga la suma de esas distancias. También calcule la suma de las distancias que hay de cada uno de los datos de la segunda distribución al mínimo de la misma. (Dé su respuesta completando la siguiente tabla.)

	Distribución 1		**Distribución 2**	
	mínimo =		mínimo =	
	valor	valor - mínimo	valor	valor - mínimo
suma de diferencias				

e. ¿Corroboran los números que encontró en el item anterior su intuición con respecto a la dispersión de las dos distribuciones que estamos analizando? Es decir, ¿se puede pensar que la suma de distancias de cada uno de los valores de la distribución al mínimo de la distribución es una buena medida de la dispersión?

f. Ahora, considere las dos distribuciones siguientes:

Distribución 3: 1, 4, 6, 9

Distribución 4: 1, 1, 2, 3

¿Cuál de las dos distribuciones es más dispersa? ¿Por qué?

g. Emplee el mismo criterio que se utilizó en el item **d.-**, para hallar números que permitan comparar la dispersión de las dos distribuciones. (Dé su respuesta completando la siguiente tabla.)

	Distribución 3		**Distribución 4**	
	mínimo =		mínimo =	
	valor	**valor - mínimo**	**valor**	**valor - mínimo**
suma de diferencias				

Al usar ese criterio, ¿se corrobora su intuición?

h. Considere la distribución:

Distribución 5: 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4

Haga el diagrama de las distribuciones 1 y 5. ¿Intuitivamente, alguna de esas dos distribuciones es más dispersa que la otra? ¿Cuál?

i. Emplee el mismo criterio que se utilizó en el item **d.-**, para comparar la dispersión de las dos distribuciones 1 y 5. ¿Se corrobora su intuición?

Observe que aunque las distribuciones tienen la misma dispersión, el criterio que veníamos utilizando y que parecía ser un buen criterio para medir y comparar la dispersión de dos conjuntos, arroja una información que en este caso no sirve, pues es contrario a la evidencia. Por tanto, la suma de todas las distancias de los datos de un conjunto al mínimo no constituye una buena herramienta para medir la dispersión pues no es *universal*.

j. Comente la validez de esta afirmación:

> La medida de dispersión definida anteriormente no sirve porque no tiene en cuenta el número de datos.

Considere entonces otro criterio para medir la dispersión: primero, **se calcula la distancia de cada uno de los datos de la distribución al mínimo de dicha distribución**; segundo, **se hace el promedio de dichas distancias**; y tercero, **se emplea ese resultado como medida de la dispersión de los datos de la correspondiente distribución.**

k. Utilice este nuevo criterio para medir la dispersión de las distribuciones 1 y 5. Y, comente la bondad de este criterio. (Dé su respuesta empleando una tabla como la siguiente.)

	Distribución 1		**Distribución 5**	
	mínimo =		mínimo =	
	valor	**valor - mínimo**	**valor**	**valor - mínimo**
suma de diferencias				
promedio de diferencias				

l. Considere las siguientes distribuciones:

Distribución 6: 1, 5, 6, 7, 8

Distribución 7: 1, 2, 3, 4, 8

¿Intuitivamente, alguna de las dos distribuciones es más dispersa que la otra? Utilice el último criterio definido, ♣*el promedio de las distancias de cada uno de los datos de la distribución al mínimo de la misma*♣ para medir la dispersión de las distribuciones 6 y 7 y compararlas. ¿Se corrobora su intuición? ¿Cuál es el problema? Emplee una tabla como la siguiente:

	Distribución 6		**Distribución 7**	
	mínimo =		mínimo =	
	valor	**valor - mínimo**	**valor**	**valor - mínimo**
suma de diferencias				
promedio de diferencias				

Debe ser claro que la última "medida de dispersión" que hemos definido tampoco es una buena herramienta para comparar la dispersión de dos distribuciones pues depende de qué tan alejado esté el mínimo del resto de los datos de la distribución. Por tanto, aún no hemos encontrado cuál es la mejor referencia con respecto a la cual debemos medir las distancias.

m. Sugiera cuál es una buena referencia con respecto a la cual se deban medir las distancias, para obtener una medida de dispersión.

Puesto que la media de una distribución es, en términos generales, un buen representante de la distribución, resulta natural pensar que la referencia que hemos estado buscando es esa medida. Bien, entonces definamos ahora como medida de dispersión **el promedio de las "distancias"** (diferencias entre los valores de la distribución y la media de la misma) **de cada uno de los datos de la distribución a la media de la distribución.**

n. Para cada una de las distribuciones 1, 5 y 6 emplee la definición dada anteriormente para medir y comparar la dispersión de los datos de cada una de las tres distribuciones.

	Ditribución 1		**Distribución 5**		**Distribución 6**	
	media =		media =		media =	
	valor	**valor - media**	**valor**	**valor - media**	**valor**	**valor - media**
suma de diferencias						
promedio de diferencias						

¿Le sorprende el resultado? ¿En qué consiste y cómo se puede resolver el problema que hemos encontrado al definir así la medida de la dispersión?

Usted debió encontrar que el problema reside en que los valores positivos (que corresponden a los datos por encima de la media) se anulan con los valores negativos (que corresponden a los datos por debajo de la media). En otras palabras, no se está haciendo el promedio de *verdaderos valores de distancia*.

o. ¿Cómo lograr que todas las diferencias sean positivas?

Como lo que nos interesa es la distancia de cada valor a la media, debemos obtener valores **positivos** (recuerde que no existen distancias negativas). Una posible manera de obtener valores positivos es elevar al cuadrado cada una de las diferencias obtenidas. (La otra forma es trabajar con el valor absoluto de las diferencias; sin embargo, no tomaremos ese camino.) Y, entonces, se puede

pensar en definir una herramienta que mida la dispersión de una distribución, como el **promedio de los cuadrados de las diferencias de cada uno de los datos a la media de la distribución.**

p. Verifique que al emplear esta última herramienta para medir y comparar la dispersión de cualquier par de distribuciones de las dadas anteriormente, el resultado que se obtiene corrobora la intuición correspondiente. Para ello compare las distribuciones 5 y 6.

	Distribución 5			**Distribución 6**		
	media =			media =		
	x_i	x_i - media	$(x_i$ - media$)^2$	x_i	x_i - media	$(x_i$ - media$)^2$
promedio de los cuadrados de las diferencias						

q. Emplee la última herramienta definida para comparar el desempeño de los dos cursos en el examen del que se habla al comienzo de esta sección. ¿Se corrobora su intuición?

Hemos encontrado, entonces, una herramienta que depende de **todos** los datos de la distribución y además tiene en cuenta el **número de datos** que hay en ella. Además, proporciona resultados que son coherentes con la observación y la intuición. Esta medida se conoce como la *varianza* de la distribución.

> La *varianza* de un conjunto de datos numéricos es una medida de su dispersión y se define como el promedio de los cuadrados de las diferencias de cada valor a la media aritmética.

r. En una cierta investigación se tomó una muestra de 10 niños y por cada uno de ellos se obtuvo una medida correspondiente a su estatura (en metros). A continuación se da la muestra de datos:

{1,25, 1,32, 1,38, 1,25, 1,32, 1,20, 1,32, 1,32, 1,25, 1,25}

Determine la estatura promedio de ese conjunto de datos. ♣*No olvide dar la respuesta en metros.*♣ Además, utilice la varianza para medir la dispersión de los datos. ¿En qué unidades se expresa esa medida? Comente ese hecho y sugiera alguna solución.

Por razones como la que usted descubrió en el caso de las estaturas, en ciertas ocasiones, el valor más comúnmente empleado para medir la dispersión es el llamado *desviación estándar* que se define como la raíz cuadrada de la varianza

> La *desviación estándar* de un conjunto de datos numéricos es una medida de su dispersión. Se define como la raíz cuadrada del promedio de los cuadrados de las distancias que hay de cada uno de los datos del conjunto a la media aritmética del mismo.

s. Para el problema planteado al iniciar este capítulo, —el que se refiere al desempeño de los alumnos de dos secciones del curso de estadística en un mismo parcial— calcule la desviación estándar de cada una de las dos distribuciones y compruebe que tales números reflejan efectivamente la diferencia de dispersión que hay entre los dos conjuntos de calificaciones.

Un resumen

Recapacitemos un poco sobre la nueva herramienta que hemos descubierto. Recordemos que el problema que da lugar a buscar esta herramienta es el hecho de que, a veces, nos encontramos con conjuntos de datos que son claramente diferentes, pero que tienen la misma medida de tendencia central. Esta diferencia se debe a que los conjuntos de datos tienen diferente dispersión. Se entiende por dispersión la cualidad que se refiere al grado de esparcimiento que hay entre los elementos de un conjunto de datos numéricos, a la variabilidad que presentan dichos datos.

En primera instancia, pensamos que el rango, que es la diferencia entre el valor máximo y el valor mínimo de los datos podría ser una herramienta adecuada para medir la dispersión. En algunos casos, esta herramienta proporciona medidas de la dispersión que están de acuerdo con la intuición. Sin embargo, también encontramos conjuntos de datos para los cuales el rango da

lugar a resultados que van en contra de la intuición. Por consiguiente, no pudimos aceptar el rango como una herramienta adecuada para medir la dispersión, puesto que deseábamos encontrar una herramienta universal.

La causa por la cual el rango no es una herramienta universal para medir la dispersión de un conjunto de datos, es que tiene en cuenta solamente dos de esos datos; y, claramente, la dispersión depende de la distribución de todos los datos. Entonces, decidimos buscar otra herramienta más general que satisficiera estas condiciones. Para ello, buscamos, primero que todo, una herramienta que tuviera en cuenta todos los datos. Y no fue difícil darnos cuenta de que esa herramienta debía medir las distancias entre los datos. El problema, entonces, era determinar con respecto a qué se medirían esas distancias. Y la respuesta fue que debían ser medidas con respecto al punto medio de los datos.

Con este concepto sencillo e intuitivo logramos encontrar nuestra nueva herramienta: ella mide las distancias de todos y cada uno de los datos con respecto a la media. Lo que siguió fue la solución de problemas prácticos al intentar expresar este concepto intuitivo en algo que nos diera como resultado un número. ¿Cómo medir las distancias? Claramente no podíamos medir las distancias como diferencias de los datos a la media, puesto que esto nos daría diferencias negativas y toda distancia debe ser mayor o igual que cero. La solución era sencilla: elevar las diferencias al cuadrado. De hecho, ese es el cuadrado de la distancia. En seguida, teníamos que encontrar una forma de resumir todo ese conjunto de cuadrados de diferencias en un sólo número. Podríamos haber pensado que bastaría con sumarlos y eso nos daría el número buscado. Sin embargo, ésta no habría sido una herramienta apropiada, puesto que entonces un conjunto de datos numeroso resultaría con una dispersión más grande que un conjunto de datos poco numeroso y no necesariamente esto debería ser así. Es por ello que decidimos dividir la suma por el número de datos, lo que es igual a sacar el promedio de los cuadrados de las diferencias. Se obtiene entonces la herramienta llamada varianza, cuya fórmula es:

$$\text{varianza} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \text{media})^2$$

La varianza es una buena medida de la dispersión de un conjunto de datos. Sin embargo, tiene un pequeño problema: puesto que considera los cuadrados de las diferencias, y no las diferencias mismas, el resultado no se encuentra en la misma escala que los datos originales. Por ejemplo, si los datos se refieren a

estaturas de personas, en centímetros, entonces la herramienta que tenemos hasta ahora nos da un resultado en centímetros cuadrados. Este problema se resuelve fácilmente: basta con sacar la raíz cuadrada del resultado que tenemos. Esta es la herramienta que hemos obtenido. Se llama desviación estándar. Si a usted le gustan las fórmulas, la que corresponde a esta herramienta se escribe:

$$\text{desv. est.} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \text{media})^2}$$

Al igual que para las medidas de tendencia central, para el caso de las medidas de dispersión, también podemos hablar de parámetro o de estimador de parámetro según que el valor se refiera a la población o a una muestra. Así, por ejemplo, si consideramos como población a P = {1, 2, 3, 4, 5} y de P se extrae la muestra M = {2, 2, 4, 5} se tiene que:

1,4142136 —la desviación estándar de P— es un parámetro

1,2990381 —la desviación estándar de M— es un estimador de la desviación estándar de P

Para distinguir los valores de los parámetros de los valores de los estimadores se emplean notaciones diferentes. En caso de que se hable de la desviación estándar de una población, el valor se denota con la letra griega σ y la varianza correspondiente se nota por σ^2. En caso de que se hable de la desviación estándar de una muestra, el valor se denota con la letra s y la varianza correspondiente se nota por s^2.

¿Cómo hallar la desviación?

Una cosa es conocer el concepto intuitivo, otra es conocer la fórmula que surge de ese concepto y, otra es tener un método para utilizar esa fórmula cuando se desee aplicar la herramienta a un conjunto de datos. Cuando obtuvimos la media como herramienta, nos encontramos con que cada uno de estos tres aspectos de la herramienta era sencillo. Sin embargo, en el caso de la

desviación estándar, aunque el concepto intuitivo es también sencillo, la herramienta es un poco más compleja cuando se expresa en una fórmula y, por consiguiente, se pueden tener dudas acerca de la manera de aplicar la fórmula para poder calcular la desviación estándar de un conjunto de datos. Pero la situación no es tan complicada, si la aproximación al problema se hace en forma ordenada. Basta observar la fórmula y seguir los pasos que ésta sugiere. En este caso, los pasos son los siguientes:

- Se hace una tabla que tenga tres columnas y tantas filas como datos haya.
- En la primera columna se escriben los datos.
- Se calcula la media de estos datos.
- En la segunda columna se escribe, frente a cada dato, la diferencia entre éste y la media.
- En la tercera columna se calcula el cuadrado de las diferencias.
- Se obtiene el promedio de los cuadrados de las diferencias.
- Se calcula la raíz cuadrada del promedio de los cuadrados de las diferencias. El número que se obtiene es la desviación estándar.

Veamos un ejemplo del método que se describió anteriormente: calcular la desviación estándar del siguiente conjunto de datos: {36, 37, 39, 40, 48}

Datos	**Diferencia con la media**	**Diferencia al cuadrado**
36	-4	16
37	-3	9
39	-1	1
40	0	0
48	8	64

Al hacer los cálculos, se obtiene que:

- La media de los datos es 40: $\bar{x} = 40$
- La media de los cuadrados de las diferencias es 18: $s^2 = 18$
- La raíz cuadrada de la media de los cuadrados de las diferencias es 4,2426: s= 4,2426
- Por consiguiente, la desviación estándar de la distribución es 4,2426.

Ejercicios

1.- Se realizó una investigación en el departamento de Santander y se tomó como referencia el municipio de Rionegro con el fin de observar cómo ha sido allí el comportamiento de la votación para Asamblea Departamental en el período comprendido entre 1970 y 1986 teniendo en cuenta la votación por partidos, y de esa manera conocer las tendencias partidistas en dicho municipio. Se obtuvieron los siguientes resultados:

Año	# votos por P. Liberal	# votos por P. Conservador
70	3.933	1.997
72	4.756	368
74	9.514	1.335
76	7.441	1.027
78	8.097	1.310
80	9.454	1.770
82	9.633	96
84	8.840	2.653
86	6.399	215

a. ¿Cuál es la población de estudio? Y, ¿cuál es la muestra de estudio?

b. ¿Cuál es la variable que se quiere medir? ¿De qué tipo es?

c. Por medio de diagramas de barras represente los datos de los votos por el Partido Liberal y por el Partido Conservador. Con base en esos diagramas determine en cuál de los dos partidos Liberal o Conservador hubo una votación más homogénea, a través del período comprendido entre 1970 y 1986.

d. Calcule dos valores que le permitan comparar la dispersión de la votación por los partidos Liberal y Conservador, a través de los 16 años que se están considerando. ¿Corroboran esas medidas la respuesta que usted dio a partir de la observación de los diagramas?

e. Si los investigadores quisieran determinar alrededor de qué dato se agrupan los resultados de la votación liberal, ¿cuál sería la medida más aconsejable para tal propósito? ¿Por qué?

f. Con base en las respuestas anteriores haga un comentario que sea pertinente para el objetivo de la investigación.

2.- Un politólogo quiso investigar los resultados de las elecciones presidenciales de 1978 con el propósito de observar y determinar qué tan dispersos fueron los datos de la votación por candidato. Su estudio tomó en cuenta algunas alcaldías menores de Bogotá del total de alcaldías de la capital. Encontró los siguientes resultados:

zonas	# de votos por Turbay A.	# de votos por Betancur C.
Chapinero	10.616	13.770
Santa Fe	6.833	7.377
San Cristóbal	7.591	8.224
Tunjuelito	8.342	7.401
Kennedy	12.521	15.092
Bosa	2.629	2.330
Teusaquillo	14.310	16.982
Mártires	6.990	5.838
A. Nariño	18.430	16.628
P. Aranda	11.741	13.489

a. Identifique la población de estudio y la muestra.

b. ¿Cuál es la variable que se está midiendo y de qué tipo es?

c. Elabore gráficas que representen la distribución de los datos.

d. Con base en la gráficas anteriores calcule las medidas que considere más adecuadas para:

- observar alrededor de que número se agrupan los datos de la votación por Belisario Betancur.
- comparar la dispersión de los datos de la votación por Betancur y por Turbay.
- ¿En cuál de las alcaldías se presentó la mayor votación por Turbay?

3.- Se conoce el potencial electoral de algunos de los municipios ubicados al sur de Bogotá. A continuación se presenta un cuadro con el potencial electoral de 15 de esos municipios:

municipio	**número de electores**
A	13.914
B	9.667
C	10.035
D	11.205
E	10.500
F	11.623
G	10.961
H	9.350
I	12.976
J	9.877
K	9.580
L	10.720
M	10.214
N	11.530
O	12.425

a. Identifique la población de estudio y la muestra.

b. ¿Cuál es la variable que se está midiendo y de qué tipo es?

c. ¿Qué medida de tendencia central considera que es la más adecuada para determinar alrededor de qué dato se agrupa el potencial electoral de dichos municipios? ¿Por qué? Determine esa medida.

d. En caso de que quisiera ver qué tan dispersos se hallan unos datos de otros, ¿qué medida aconsejaría usar? ¿Por qué?

También se conoce el número de personas que votan en cada uno de esos 15 municipios:

municipio	número de votantes
A	9.554
B	8.721
C	9.720
D	10.442
E	9.530
F	9.560
G	8.925
H	8.023
I	10.330
J	8.905
K	8.030
L	9.932
M	9.555
N	10.024
O	9.327

e. En este caso, ¿cuál es la variable de estudio y de qué tipo es?

f. Realice una gráfica con los datos de potencial electoral y número de votantes para observar el comportamiento de dichas variables y de esa manera determinar el nivel de abstención de cada municipio.

g. Si usted como investigador buscara determinar alrededor de qué dato se agrupa el potencial electoral y alrededor de qué dato se agrupa el número de votantes, ¿qué medidas hallaría? Y si quisiera comparar la dispersión de esos dos conjuntos de datos, ¿qué medidas hallaría? Establezca esa comparación.

4.- En el colegio X de Bogotá, se está realizando un experimento para comprobar si realmente la enseñanza de las matemáticas por computador supera la enseñanza tradicional, en cuanto a resultados inmediatos de los niños. Dicho experimento se realizó en los estudiantes de cuarto de primaria y ocurrió así: se tomó un *grupo control* formado por 17 de esos niños, elegidos de su curso aleatoriamente, se les enseñó un cierto tema a la manera tradicional. También se tomó un *grupo experimental* formado por otros 17 niños, elegidos del mismo curso en forma aleatoria y se les enseñó el mismo tema con asistencia del computador. Cuando terminaron de ver el tema, se les hizo a los 34 alumnos el mismo examen. Las calificaciones obtenidas por los niños se presentan a continuación.

Gpo. control	Gpo. experimental	Gpo. control	Gpo. experimental
38	40	40	40
35	38	33	35
36	35	39	42
41	45	40	44
42	45	38	38
36	40	40	40
32	35	30	38
29	39	41	40
28	38		

a. Identifique la población de estudio y la muestra a partir de la cual se van a realizar las inferencias.

b. ¿Cuál es la variable que se está midiendo?

c. Haga una tabla de frecuencias agrupadas para cada uno de los casos, es decir, para el grupo control y para el grupo experimental.

d. Haga dos diagramas que representen los datos obtenidos en los dos casos.

e. ¿Cuál de las medidas de tendencia central podría describir el comportamiento de la variable en cada una de los dos grupos?

f. Compare la dispersión de las calificaciones en cada uno de los dos grupos.

g. ¿Cree usted que las calificaciones obtenidas en los dos grupos corroboran la hipótesis de que la enseñanza por computador supera en resultados inmediatos a la enseñanza tradicional? Explique su respuesta.

h. Independientemente de los resultados, ¿cree usted que es ventajosa la enseñanza por computador? Justifique su respuesta.

5.- Un estudio sociolingüístico desarrollado en una universidad pública de Bogotá en 1990 pretende demostrar la hipótesis, anteriormente trabajada por lingüistas norteamericanos, de que existe una relación entre la pertenencia a un grupo social y la preferencia por el "lenguaje, acento o dialecto distintivo del grupo social al que desearía pertenecer".[17] El estudio pedía a diez sujetos de estrato socioeconómico medio bajo evaluar de 1 a 10 el lenguaje de tres personas cuyas voces se encontraban registradas en una cinta magnetofónica. La primera persona era de clase alta; la segunda de clase media alta; y la tercera de clase baja. Las calificaciones que dieron los sujetos a cada una de las tres grabaciones fueron:

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

a. ¿Cuál es el problema de estudio?

b. ¿Qué variables se toman en consideración? ¿De qué tipo son?

c. Mirando rápidamente los datos presentados, ¿cuál de los tres grupos de calificaciones está más disperso?

d. Calcule la varianza de cada grupo de calificaciones. ¿Confirman estos valores su respuesta a la pregunta anterior? Justifique.

17 Tucker, Richard and Wallace Lambert. "White and Negro Listener's Reactions to Various American-English Dialects". *Advances in the Sociology of Language, Joshua Fishman, ed. Paris: Mouton Publishers, 1972, p. 176.*

e. Para el primer grupo de calificaciones, ¿qué porcentaje de observaciones se alejan de la media más de dos desviaciones estándar a derecha e izquierda.

f. Teniendo en cuenta las calificaciones a la persona 2 y a la persona 3, suponga que un sujeto calificó con 9,2 ambas personas. ¿En cuál de los dos grupos esa calificación es más "rara"? (Piense en qué tan alejado se encuentra el 9,2 de la media de cada uno de los grupos de calificaciones.)

g. "Los sujetos de una clase social quieren ascender lenta y paulatinamente en la estratificación socioeconómica y no hacerlo rápida y bruscamente". Valiéndose de la información que posee para los diferentes niveles socioeconómicos pruebe esta hipótesis justificando sus afirmaciones.

6.- Considere el siguiente diálogo:

Askanio: Ana Liza, ¿qué calificaciones llevas en los parciales del curso de estadística?

Ana Liza: En el primero saqué 3,3, en el segundo, 3,5, y en el tercero, 3,9.

Askanio: ¡Uf! Vas muy bien. Seguro que en el cuarto sacas por encima de 4,0.

Ana Liza: Sí me ha ido bien. Sin embargo, con respecto al curso, en el primer parcial me fue mejor que en el segundo y en el segundo me fue mejor que en el tercero.

Eduardo: **(hermano de Askanio)** ¿Cómo es esa locura? ¿Te fue mejor en el que sacaste 3,3 que en el que sacaste 3,9?

Ana Liza: Así es, Eduardo. Una calificación en sí misma no tiene significado. Pero se llena de significado cuando se compara con otras calificaciones.

Eduardo: Entiendo. Entonces tiene que ser que en el primer parcial, en general, le fue mal al grupo. Y, además que poco a poco ha ido mejorando.

a. Con respecto al diálogo anterior, demuestre que la segunda afirmación de Ana Liza es verdadera. Para ello utilice la siguiente información:

# de parcial	media	desviación estándar
1	2,,8	0,,56
2	3,,0	0,,60
3	3,,5	0,,70

b. Estrella y Ana Liza son compañeras en el curso de estadística. Si en el segundo parcial, la calificación de Estrella está a 2,5 desviaciones estándar por encima de la media, ¿qué calificación obtuvo Estrella?

7.- Considere el lanzamiento de un par de dados durante 3.600 veces. Cada vez se registra el valor de la suma de los resultados de las caras superiores. Los resultados se presentan en la siguiente tabla:

Valor suma	2	3	4	5	6	7	8	9	10	11	12
Frecuencia	100	200	300	400	500	600	500	400	300	200	100

a. Describa de la manera más completa posible la distribución presentada anteriormente. (No recurra a la enumeración de los datos.)

b. ¿Cree usted que la distribución que se presentó se obtuvo empíricamente? Justifique su respuesta.

c. ¿En qué porcentaje del total de lanzamientos se obtuvo un valor de la suma que estuviera a menos de 1 desviación estándar de la media?

d. ¿En qué porcentaje del total de lanzamientos se obtuvo un valor de la suma que estuviera a menos de 2 desviaciones estándar de la media?

e. ¿En qué porcentaje del total de lanzamientos se obtuvo un valor de la suma que estuviera a menos de 3 desviaciones estándar de la media?

8.- Considere las tres distribuciones que se presentan a continuación:

Distribución 1						
Observación	1	2	3	4	5	6
Frecuencia	100	100	100	100	100	100

Distribución 2						
Observación	1	1,,5	2	2,,5	3	3,,5
Frecuencia	100	100	100	100	100	100
Distribución 3						
Observación	1	4/3	5/3	2	7/3	8/3
Frecuencia	100	100	100	100	100	100

a. ¿Cuál de las tres distribuciones presentadas es la más dispersa, y cuál la menos dispersa? Justifique matemáticamente su respuesta.

b. Para cada una de las tres distribuciones determine el porcentaje del total de observaciones que distan menos de 1 vez la desviación estándar correspondiente.

c. Para cada una de las tres distribuciones determine el porcentaje del total de observaciones que distan menos de 2 veces la desviación estándar correspondiente.

d. ¿Le sorprenden los resultados obtenidos en los dos items anteriores? Explique su respuesta.

9.- Invéntese una distribución de frecuencias (que refleje el comportamiento de una determinada variable, es decir, invéntese el conjunto de datos pensando en una situación real) que usted considere muy dispersa.

a. Para el conjunto de datos que ha dado, determine qué porcentaje del total de observaciones está a menos de 1 desviación estándar de la media.

b. Determine también el porcentaje de observaciones que está a menos de 2 desviaciones estándar de la media.

c. Determine a cuántas desviaciones estándar de la media quedan contempladas todas las observaciones del conjunto de datos que usted dio.

10.- Con base en las respuestas que usted ha encontrado a las últimas cuatro preguntas, intente explicar el papel que desempeña la desviación estándar al hablar de distribuciones de datos.

La ley

Lectura[18]

por Robert M. Coates

El indicio de que las cosas estaban saliéndose de su cauce normal vino una tarde de finales de la década de 1940. Simplemente lo que pasó fue que entre las siete y las nueve de aquella tarde el puente Triborouhg[19] tuvo la concentración de tráfico saliente más elevada de su historia.

Esto era raro, porque se trataba de la noche de un día laborable (para ser precisos, un miércoles) y aunque el tiempo era agradablemente benigno y claro, con una luna que estaba lo bastante crecida para atraer un buen número de motoristas a abandonar la ciudad, estos hechos por sí solos no eran suficientes para explicar el extraño fenómeno. Las dos noches precedentes, aunque fueron igualmente tranquilas e iluminadas, no provocaron en ningún puente o carretera un fenómeno semejante.

Por de pronto, el personal del puente fue cogido por sorpresa. Una gran arteria de tráfico como el Triborough, opera en condiciones normalmente previsibles. Todo el tráfico rodado, como la mayoría de actividades humanas que se realizan en gran escala, obedece a la Ley de los Promedios, esta grandiosa y vieja regla, que establece que las acciones de la gente en grandes ciudades siguen siempre modelos estables; basándose en la experiencia pasada, siempre había sido posible predecir, con toda exactitud, el número de coches que cruzaría el puente a una hora determinada del día o de la noche. En esta ocasión todas las reglas fallaron.

Las horas que transcurren desde las siete hasta cerca de medianoche son normalmente tranquilas en el puente. Pero aquella noche parecía como si todos los motoristas o buena parte de ellos se hubieran puesto de acuerdo para rom-

18 Tomada de *Sigma*, Grijalbo, Vol. 6, pp. 205-208.

19 Es uno de los puentes que une a Manhattan con Nueva Jersey.

per la tradición. Empezando casi exactamente a las siete, los coches se dirigieron hacia el puente en tal número y con tal rapidez, que los empleados de las taquillas se vieron desbordados por el trabajo, casi desde el principio. Pronto se vio que no era una congestión momentánea, y cuando se hizo evidente que el tráfico prometía adquirir proporciones gigantescas, se trasladaron a toda prisa policías hacia el lugar del suceso.

Los coches fluían de todas direcciones, de la ruta de Bronx y de la de Manhattan, de la Calle 125 y de East River Drive. (En un extremo de la aglomeración, apretada línea de luces de coches que se perdía de vista hacia el sur de la calle 89, al mismo tiempo que la aglomeración cruzaba la ciudad de Manhattan interrumpiendo el tráfico hacia el oeste de la avenida Amsterdam.) Quizá lo más sorprendente de esta manifestación era el hecho de que parecía no tener ninguna causa plausible.

De vez en cuando, mientras los guardias de la taquilla del peaje atendían el aparentemente infinito río de coches, preguntaban a sus ocupantes; pronto se vio claramente que los mismos participantes de la monstruosa obstrucción eran tan ignorantes de las razones que la había ocasionado como ajenos a ella. El sargento Alfonse O'Toole, que mandaba el destacamento encargado de la carretera de Bronx, hizo un informe muy significativo. "Les hice algunas preguntas", dijo "¿Es que hay algún partido de fútbol del que no tengamos conocimiento? O, ¿quizás se trata de carreras de caballos?" Pero lo más divertido era que todos me preguntaban: "¿Qué es este gentío, Mac?" Y yo solamente les miraba. Me acuerdo que había un muchacho con una chica al lado de un Ford convertible, y cuando me hizo esta pregunta le respondí "¿Estás en medio de la multitud, no es verdad? ¿Qué te ha traído aquí?" le pregunté. Y el chico, mirándome, dijo: "¿Yo?, tan sólo he venido a dar un paseo a la luz de la luna. Pero si hubiera sabido que había una aglomeración así..." dijo. Y entonces me pregunté: "¿Hay algún lugar para que pueda dar la vuelta y salir de aquí?" A la mañana siguiente el *Herald Tribune* relató este suceso, "parecía como si todos los propietarios de coches de Manhattan hubieran decidido aquella noche dirigirse hacia Long Island".

El incidente era tan extraordinario que ocupó la primera plana de todos los periódicos a la mañana siguiente, y a causa de ello, muchos sucesos parecidos, que de otra forma no hubieran sido nunca remarcados, fueron extensamente comentados. Así, el propietario del teatro Aramis, en la Octava Avenida, explicó que mientras durante algunos días su sala había permanecido prácticamente vacía, otros se había llenado hasta los topes. Los propietarios de Luncheon notaron que con el aumento de clientes estaban desarrollando más la

costumbre de hacer operaciones con artículos específicos; un día todo el mundo pedía paletillas de ternera asada con salsa, mientras que otro todos pedían panecillos de viena, y al cordero asado nadie le hacía caso. Un hombre que dirigía un pequeño almacén de baratijas en Bayside explicó que entraron en su tienda en el espacio de cuatro días, 274 clientes pidiendo un ovillo de hilo rosa.

En un período de normalidad, esta noticia se hubiera escrito en los periódicos o bien como relleno, o bien en la sección de curiosidades; sin embargo, en la situación actual, adquirían una mayor relevancia. Finalmente, se hizo evidente que algo extraño estaba sucediendo con las costumbres de la gente, las cuales estaban sufriendo un cambio tan radical e imprevisto, como lo que sucede cuando en una excursión en barco todos los pasajeros a la vez se inclinan hacia un lado u otro de la embarcación. Sucedió en un día de diciembre que, casi increíblemente, por primera vez la Twentieth Century Limited salió del puerto de Nueva York en dirección de Chicago con sólo tres pasajeros a bordo; fue entonces cuando los empresarios se dieron cuenta de las desastrosas consecuencias que podía traer el nuevo curso de las cosas.

Hasta entonces, por ejemplo, la Central de Nueva York podía actuar con cierta confianza bajo el supuesto de que hubiera unas mil personas en Nueva York que tuvieran relaciones comerciales con Chicago, y que en cualquier día laboral algunos cientos de ellas tendrían que ir allí. El empresario teatral estaba seguro de que el número de clientes en cada función se regulara por sí mismo, y que aproximadamente habría el mismo número de personas que deseara ver la obra del jueves que las que había habido el martes o el miércoles. Ahora ya no se podía estar seguro de nada. La Ley de los Promedios había sido tirada por la borda, y si el efecto que ello tendría en los negocios prometía ser desastroso, no iba a ser menos para los consumidores.

Por ejemplo, cuando una señora se dirigía a la ciudad para ir de compras, nunca podía estar segura si en el almacén Macy iba a encontrar una avalancha de gente, clientes de otros tenderos, o un desierto vacío con resonantes pasillos y dependientes cruzados de brazos. Y cuando los individuos veíanse obligados a tomar alguna decisión, se producía una extraña incertidumbre. La gente se preguntaba a sí misma: "¿Puedo hacer esto o no?", sabiendo que si lo *dejaban de hacer*, perderían la más satisfactoria posibilidad entre todas las posibilidades de poseer Jones Beach. Los negocios languidecían y una especie de desesperada incertidumbre flotaba sobre todo el mundo.

Cuando la situación resultó tan grave, fue inevitable que se llamara al Congreso para que éste tomara alguna decisión al respecto. En efecto se convocó el Congreso, y debe decirse que su actuación fue magnífica. Se nombró un Comité, representado por Houses y dirigido por el Senador J. Wing Looper (R), de Indiana; y después de una exhaustiva investigación, el Comité se vio obligado a concluir que no había evidencia alguna de que existiera instigación comunista. Era evidente que los trastornos fueron ocasionados por la conducta fortuita de la gente. El problema estaba en encontrar una solución. No se puede procesar a una nación entera, sobre todo en materia tan vana como ésta. El Senador Slooper señaló audazmente: "Ustedes pueden controlarlo", y finalmente se aprobó un sistema de reeducación y reforma, encaminado a conducir otra vez a la gente —citamos textualmente al Senador Slooper— "a las regularidades fundamentales, a los promedios sencillos en la forma de vida americana".

En el curso de las investigaciones realizadas por el Comité se descubrió, con el consiguiente asombro de todo el mundo, que la Ley de los Promedios nunca había sido incluida en el Cuerpo Doctrinal de la jurisprudencia federal, y aunque los defensores de los States Rights se rebelaron violentamente, el olvido fue corregido, por una enmienda constitucional y por una ley —la Hills-Slooter Act— que la complementaba. De acuerdo con el contenido del Act, se *obligaba* a la gente a estar proporcionados, y la forma más fácil de asegurar que ello se cumpliría era hacer una división en el alfabeto, por la que se permitiría actuar a cada individuo sólo cuando le correspondiera según su apellido. De este modo una persona cuyo nombre empezara con "G", "N", o "U", por ejemplo sólo podría ir al teatro el martes, y a los partidos de fútbol sólo los jueves, mientras que sus visitas al supermercado deberían hacerse entre las diez y las doce del mediodía de los lunes.

Desde luego la Ley tuvo sus inconvenientes. Tuvo un efecto desequilibrador en las funciones de teatro y en otras actividades sociales, y el coste de hacer cumplir a la gente las normas fue increíblemente pesado. Al final hubo que hacer tantas reformas —tal como la que permitía a los caballeros ir a todas las funciones o actos sociales acompañados de su novia, independientemente de la letra con que empezara su apellido—, que los tribunales se encontraron sin saber qué hacer ante las violaciones de la ley.

Por este lado, sin embargo, la ley servía a sus propósitos, porque inducía —bastante mecánicamente, es verdad, pero también de forma adecuada— a volver a aquella vida promediada que le gustaba al Senador Slooper. Todo habría ido bien si no fuera porque, después de un año o un poco más, informes

inquietantes empezaron a llegar desde las regiones más apartadas del país. Parecía que una extraña ola de prosperidad había invadido aquellos lugares, que anteriormente se habían considerado áreas marginales. Los montañeros de Tenessee estaban comprando Packards convertibles, y Sears Roebuck explicó que en el Ozarks las ventas de artículos de lujo se habían incrementado en un novecientos por cien. En las miserables regiones de Vermont, hombres que con muchos apuros podían vivir de lo que la tierra les daba, ahora enviaban a sus hijo a estudiar a Europa y encargaban buenos cigarros de Nueva York. Parecía como si la Ley de las Rentas Disminuyentes también se comportara de forma loca.

a. Según la lectura, una consecuencia de la ley de los promedios es que:

> "Las acciones de la gente en grandes ciudades siguen siempre modelos estables."

¿Está usted de acuerdo con esa afirmación? Si así es, dé cuatro ejemplos de comportamientos humanos en los que sea evidente el cumplimiento de tal ley. Si usted no está de acuerdo con lo que expone la ley, justifique su posición.

b. Comente la siguiente afirmación que aparece en la lectura:

> "(...) basándose en la experiencia pasada, siempre había sido posible predecir, con toda exactitud, el número de coches que cruzaría el puente a una hora determinada del día o de la noche."

c. Si se toma como marco de referencia el conjunto de los estudiantes de la universidad, ♣ *los que fueron, los que son y los que serán* ♣ determine cuáles de las siguientes variables se comportan de acuerdo a la ley de los promedios. Además, para aquellas variables que siguen algún modelo estable, haga una gráfica que represente tal modelo.

- La estatura de hombres y mujeres
- La estatura de mujeres

- El número de semestres cursados por un estudiante regular, hasta terminar su carrera
- El valor de la matrícula que pagó un estudiante este semestre
- Horas del día en que un estudiante recibe clase
- Carrera
- Lugar de procedencia
- Distancia del lugar de residencia de un estudiante a la universidad

d. En el mismo marco de referencia del ejercicio anterior, determine cuatro variables que se comporten en forma similar a la variable estatura de las mujeres. ♣ *No tienen que asumir los mismos valores, sino las mismas características fundamentales.*♣

Distribución normal

Introducción

La organización y el resumen de datos, las medidas de tendencia central y las medidas de dispersión —que son los temas de estadística tratados hasta ahora— son herramientas útiles, que es importante saber emplear adecuadamente, en especial para poder describir una muestra.

Sin embargo, si se quiere abordar y solucionar un problema desde el punto de vista de la estadística, con frecuencia el trabajo no se limita a tomar una muestra y describirla. La gran contribución de la estadística va más allá de la descripción de una o más muestras. Tiene que ver con la población: permite conocer, con algún grado de certidumbre, características de las poblaciones (de las cuales provienen las muestras con base en las cuales se trabaja) que no se pueden conocer de manera directa porque dichas poblaciones son infinitas o tan grandes y complejas que se hace imposible abarcarlas totalmente en un estudio.

En capítulos posteriores se presentan las metodologías estadísticas que permiten obtener información de la población a partir de la información de una muestra. En este capítulo nos interesa descubrir la existencia de una distribución con la cual se pueden modelar diversas situaciones tanto reales como teóricas que tienen que ver con el comportamiento de una variable cuantitativa continua. El modelo del que se está hablando es el de la *distribución normal estándar*; no es el único modelo estadístico del que se dispone pero es quizás el más importante puesto que es la base para comprender los procedimientos que involucran otros modelos, y además, sus características sirven de fundamento para las inferencias que se realizarán posteriormente.

Motivación

Considere los siguientes problemas:

Problema 1. ¿Cuál es el porcentaje de alumnos de su universidad cuyo promedio ponderado es inferior al suyo?

Problema 2. ¿Cuál es la estatura mínima que debe tener un estudiante varón de su universidad para poder pertenecer al equipo de baloncesto, si se quiere que quienes conformen el equipo tengan una estatura superior a la del 90% de todos los estudiantes varones de la universidad?

a. Si usted quisiera responder al problema 1, ¿qué datos tendría que conocer? Suponga que conoce la información necesaria; enuncie, entonces, los pasos que daría para solucionar el problema.

b. Si usted quisiera responder al problema 2, ¿qué datos tendría que conocer? Suponga que conoce la información necesaria; enuncie, entonces, los pasos que daría para solucionar el problema.

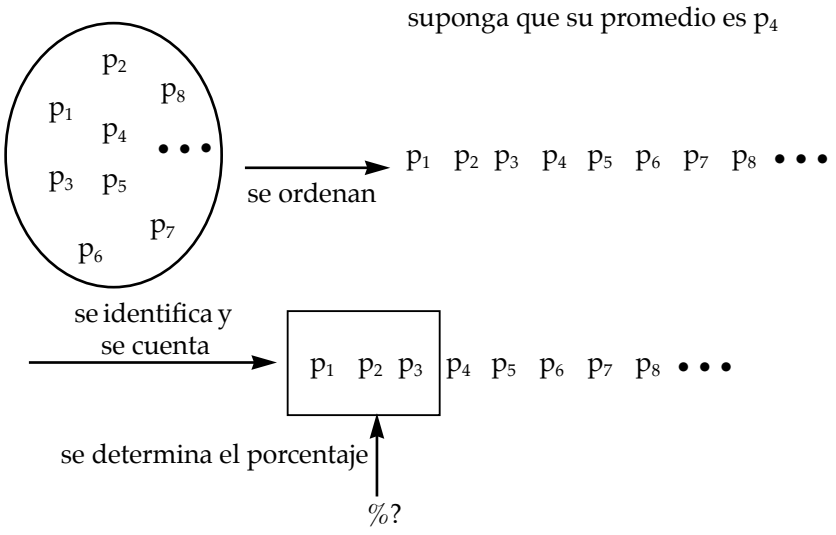
En la respuesta que dio a la pregunta **a.**, usted debió enumerar una serie de datos que se requieren para solucionar el problema 1:

- su promedio ponderado
- el promedio ponderado de todos los estudiantes de la universidad
- el número total de estudiantes

Además, debió enunciar como pasos que se deben seguir para dar solución al problema planteado, los siguientes:

- **Ordenar** los promedios ponderados (por ejemplo, en orden ascendente).
- En esa distribución de los datos, **identificar** el valor particular del promedio ponderado que está haciendo de referencia, (su promedio ponderado).
- **Contar** cuántos promedios son menores que la referencia.
- **Establecer qué porcentaje del total de promedios**, es el número encontrado anteriormente. (Y, ese porcentaje es la respuesta al problema.)

El siguiente esquema puede resumir el proceso a seguir:



Ahora vamos a proponer una situación concreta e hipotética, en la cual se lleve a la práctica lo dicho. Suponga que en la universidad hay tan sólo 50 alumnos y los promedios ponderados de ellos son:

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

Suponga que el problema para responder es:

> Si su promedio ponderado es 3,6, ¿cuál es el porcentaje de alumnos de su universidad, cuyo promedio ponderado es inferior al suyo?

c. Elabore una tabla de distribución de frecuencias de tales promedios.

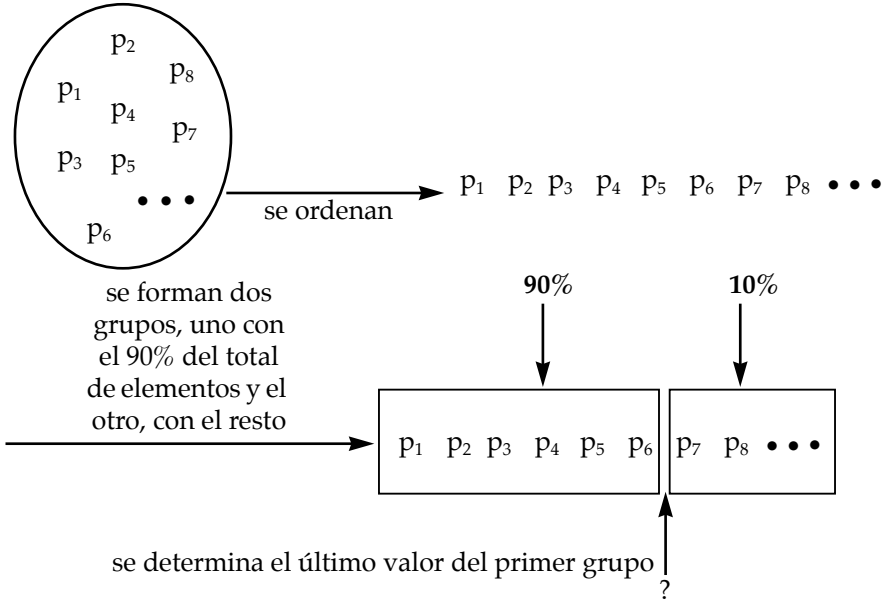
d. Muestre que el 18% del total de alumnos tiene promedio ponderado inferior a 3,6.

En la respuesta que dio a la pregunta **b.**, usted debió enumerar una serie de datos que se requieren para solucionar el problema 2:

- la estatura de todos los estudiantes varones de la universidad
- el número total de estudiantes varones

Además, debió enunciar como pasos que se deben seguir para dar solución al problema planteado, los siguientes:

- **Ordenar** las estaturas de los estudiantes varones (por ejemplo, en orden ascendente).
- **Determinar cuántas de las estaturas** constituyen el 90% del total de estaturas.
- En la distribución realizada anteriormente, **contar** tantas estaturas como lo indique el número que corresponde al 90% del total de estaturas.
- **Identificar el valor de la estatura**, que divide al grupo de estaturas en dos: uno, que contiene al 90% del total y el otro, el que contiene los otros valores de estaturas. (Y, el valor de esa estatura es la respuesta al problema.)



El esquema anterior resume los pasos a seguir.

Ahora vamos a proponer una situación concreta e hipotética, en la cual se lleve a la práctica lo dicho. Suponga que en la universidad hay tan sólo 30 alumnos varones y que sus estaturas —medidas en metros— son:

1,74	1,78	1,72	1,70	1,72	1,68	1,70	1,65	1,74	1,70
1,72	1,72	1,70	1,68	1,74	1,68	1,72	1,70	1,72	1,74
1,76	1,74	1,70	1,72	1,76	1,65	1,72	1,74	1,74	1,72

Entonces, el problema que tiene que responder es:

> ¿Cuál es la estatura que debe tener un estudiante varón de la universidad para poder afirmar que dicha estatura es superior a la estatura del 90% de todos los estudiantes varones de la universidad?

e. Elabore una tabla de distribución de frecuencias de las estaturas.

f. Muestre que 1,75 es el valor que supera el 90% de los valores del grupo total.

Ahora, para usted debe ser claro que la solución a los dos problemas planteados al inicio de la sección, **no** es difícil en sí misma. Sin embargo, si se pretende solucionar los problemas como usted lo sugirió en **a.** y en **b.**, hay factores que dificultan de manera parcial o total el encontrar la solución: en la mayoría de los casos, hay una gran cantidad de información involucrada y no es posible conocerla toda y aun en el caso de que se conozca toda la información, el volumen de ella es tal, que los procesos sugeridos por usted exigen un trabajo largo y tedioso que se puede evitar.

Por otro lado, también ha de ser claro para usted que problemas como los dos que se enuncian, no necesariamente son artificiales; es decir, existen situaciones concretas de interés particular y general que conducen al planteamiento de enunciados similares y, por tanto, surge la necesidad de encontrar una manera eficiente y ágil de abordarlos y de solucionarlos. Pues bien, en este capítulo vamos a descubrir un modelo que nos permita responder a preguntas de tal estilo.

En busca de un modelo

Se quiere construir un modelo que se pueda emplear para responder preguntas acerca del comportamiento de variables tales como la estatura, el peso, la presión arterial, el tiempo de duración de ciertos procesos biológicos, el promedio ponderado, etc. Por tanto, es natural estudiar las características de tales variables, de modo que ellas estén presentes en el modelo que se construya. Vamos pues a realizar una breve reflexión sobre tales características.

Sobre la estatura

Para responder las siguientes preguntas, usted debe emplear solamente su intuición, su sentido común, y su observación. ♣ *No se le pide que haga cálculos.*♣ Además, considere como población en la cual va a observar el comportamiento de la variable estatura, el conjunto de todos los estudiantes varones de la universidad y suponga que este conjunto es muy numeroso.

a. ¿Qué tipo de variable es la variable estatura? Explique su respuesta.

b. ¿Qué valores puede tomar la variable? En otras palabras, dé el intervalo de variación de la variable.

c. ¿Existe algún valor para la estatura, alrededor del cual se agrupe la mayor parte de las observaciones? ¿Cuál es?

d. ¿Cuál es la estatura promedio?

e. ¿Qué ocurre con la frecuencia de los valores de las estaturas a medida que éstos se alejan de la estatura media?

f. ¿Presenta la gráfica de la variable, alguna simetría? Explique.

g. ¿Qué porcentaje de la población debe tener estatura inferior a la estatura promedio? Y, ¿superior a la estatura promedio?

h. ¿Qué relación de orden existe entre la estatura promedio, la estatura moda y la estatura mediana? Explique su respuesta.

i. ¿Qué tan dispersa es la distribución? Determine el rango de la estatura. ¿Cuántas veces, aproximadamente, cree usted que "cabe" la desviación estándar en ese rango?

j. Haga un bosquejo de la gráfica que representa el comportamiento de la variable en cuestión.

Sobre el promedio ponderado y otras variables

a. Tenga en cuenta los aspectos mencionados en las preguntas sobre la variable estatura y con base en ellos represente gráficamente el comportamiento de cada una de las siguientes variables:[20]

- *Promedio ponderado* de los estudiantes de segundo semestre que estudian la misma carrera que usted y en la misma universidad.
- *Peso* de una alumna de 20 años que estudia en la universidad en la que usted estudia.
- *Presión arterial de un varón de 60 años.*

b. Compare entre sí las gráficas que elaboró para ilustrar el comportamiento de cada una de las cuatro variables: estatura, promedio ponderado, peso y presión arterial. ¿Existen diferencias esenciales en la forma? ¿Le sorprende ese resultado?

Usted debió encontrar que existe similitud en la forma como se distribuyen los valores de las cuatro variables que se están considerando. Por tanto, tiene sentido pensar que el modelo que estamos buscando sí resultará útil para dar cuenta del comportamiento de muchas variables.

Ahora veamos si su intuición y su observación de la realidad son correctos en el caso de las estaturas. En este punto se quiere que usted pueda observar especialmente dos aspectos, a saber:

- Cómo cambia la forma de la distribución de la variable a medida que se aumenta **considerablemente** la cantidad de información
- Cuál es la forma de la distribución de la variable en la población

20 En cada caso, suponga que la población en la que se está analizando el comportamiento de la variable es muy numerosa.

Vamos a suponer que en la universidad hay 3.600 estudiantes varones y para ver los dos aspectos mencionados anteriormente realizaremos lo siguiente:

- En primer lugar, se tomará una muestra aleatoria de la población, de tamaño 40 y se hará la distribución de la variable para esa muestra.
- En segundo lugar, se tomará de la población, una segunda muestra aleatoria, también de tamaño 40, que no incluya a ninguna de las personas de la primera muestra; y la información que dé esa muestra con respecto a la estatura, se adicionará a la información de la primera muestra; es decir, es como si ahora se tuviera una muestra de tamaño 80. Para esa muestra se hará la distribución de la variable.
- En tercer lugar, se tomará otra muestra aleatoria de la población, de tamaño 120, (todos los sujetos de esa muestra serán diferentes a los que conformaron las dos primeras muestras). La información que arroje esa muestra se adicionará a la información correspondiente a la unión de las dos primeras muestras. Es entonces, como si se hubiera tomado una muestra de tamaño 200. Para esa muestra se hará la distribución de los valores de la variable.

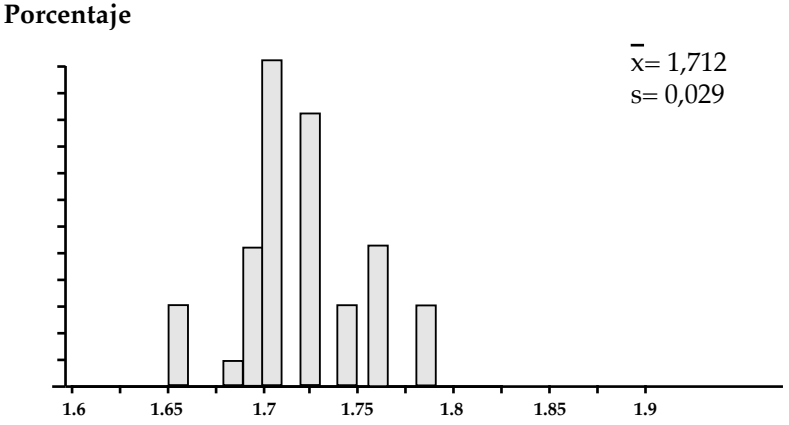
Comparemos formas

De acuerdo con lo dicho anteriormente, la información que suministra la primera muestra, de tamaño 40, está dada en la siguiente tabla (medidas en metros, aproximadas hasta centímetros).

1,70	1,72	1,74	1,65	1,72	1,70	1,70	1,72	1,70	1,72
1,78	1,65	1,70	1,72	1,70	1,68	1,65	1,69	1,70	1,74
1,72	1,76	1,76	1,69	1,76	1,72	1,69	1,72	1,69	1,70
1,69	1,70	1,72	1,70	1,74	1,76	1,70	1,76	1,70	1,72

a. Haga una tabla de distribución de frecuencias de la variable estatura en la muestra de tamaño 40.

A continuación se presenta un diagrama de la distribución de la estatura en esa muestra. (Observe que hay valores de la estatura cuya frecuencia es 0.)



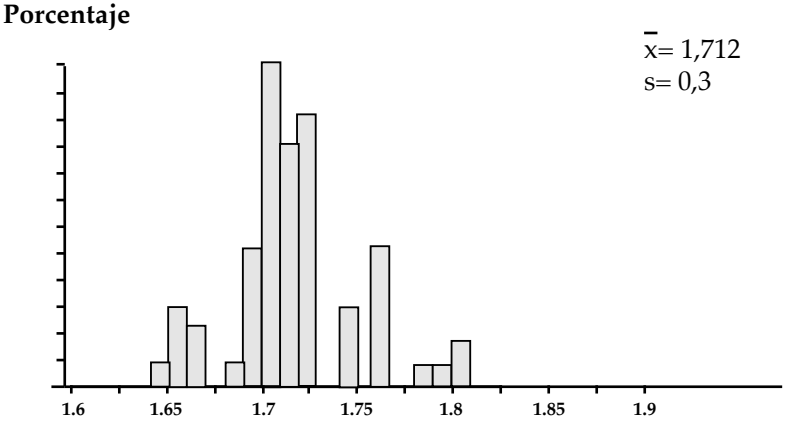
Estatura (metros) en la muestra de tamaño 40

La información suministrada por la segunda muestra, de tamaño 40, aparece a continuación:

1,71	1,76	1,79	1,80	1,71	1,70	1,74	1,72	1,76	1,71
1,69	1,72	1,71	1,69	1,76	1,68	1,66	1,71	1,70	1,76
1,80	1,71	1,74	1,72	1,66	1,71	1,71	1,76	1,68	1,74
1,71	1,70	1,66	1,68	1,71	1,64	1,68	1,65	1,71	1,69

b. Acumule la información de esta muestra a la de la primera muestra, con el fin de tener la muestra de tamaño 80. Haga una tabla de distribución de frecuencias de la variable estatura en la muestra de tamaño 80.

El siguiente diagrama muestra la distribución de la variable en la muestra de tamaño 80.



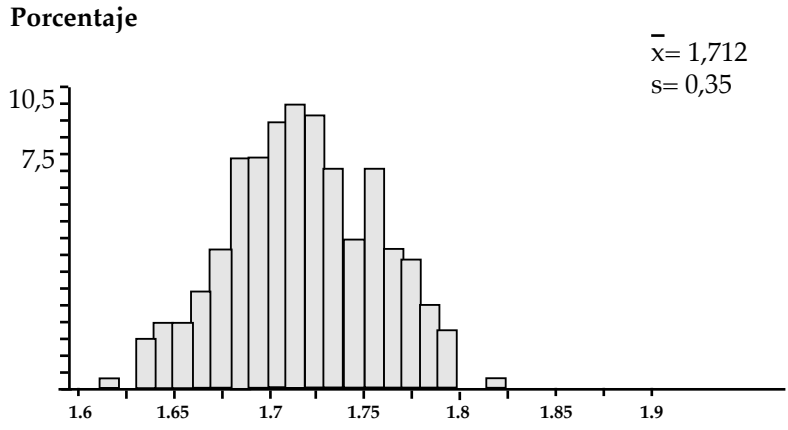
Estatura (metros) en la muestra de tamaño 80

En la última muestra tomada, de tamaño 120, la información es:

Estatura (m)	Frecuencia	Estatura (m)	Frecuencia
1,62	1	1,73	20
1,64	3	1,74	9
1,65	1	1,75	12
1,66	2	1,76	5
1,67	8	1,77	11
1,68	6	1,78	9
1,69	8	1,79	5
1,70	1	1,80	2
1,71	8	1,82	1
1,72	8		

c. Adicione la información de esta última muestra con la de la muestra de tamaño 80 y haga la correspondiente tabla de distribución de frecuencias de la variable.

El diagrama correspondiente a este caso se presenta a continuación:



Estatura (metros) en la muestra de tamaño 200

d. Compare las gráficas de las tres distribuciones, que se han obtenido a medida que se aumenta la cantidad de información, según los siguientes aspectos:

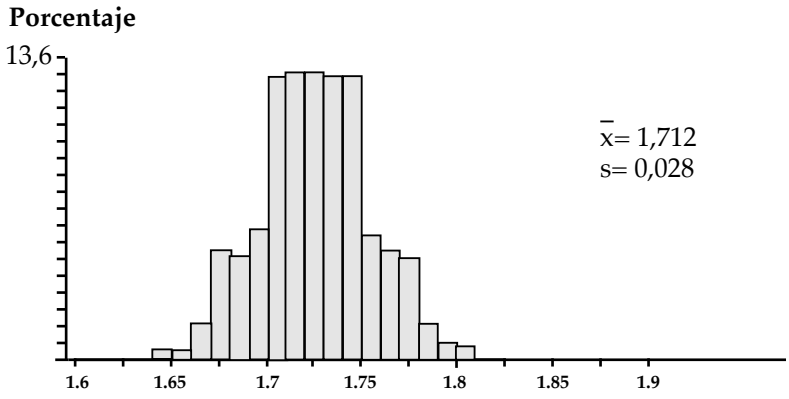
- número de picos (forma de la curva que une los puntos más altos de las columnas)
- localización del valor de la estatura alrededor del cual se agrupan la mayoría de los valores
- simetría de la curva
- frecuencia de los valores más alejados de la estatura promedio

Aunque en la mayoría de los casos no se tiene acceso a la información de toda la población y por tanto no se puede conocer con absoluta certeza la forma de la distribución de la variable en la población, en este caso, vamos a suponer que sí conocemos las estaturas de todos los estudiantes varones de la universidad. A continuación se da una tabla de frecuencias:

Estatura (m)	Frecuencia	Estatura (m)	Frecuencia
1,62	1	1,73	487
1,64	18	1,74	489
1,65	17	1,75	182

Estatura (m)	Frecuencia	Estatura (m)	Frecuencia
1,66	50	1,76	162
1,67	156	1,77	148
1,68	148	1,78	49
1,69	188	1,79	23
1,70	485	1,80	17
1,71	488	1,82	1
1,72	491		

El siguiente diagrama representa gráficamente el comportamiento de la variable estatura en la población.



Estatura (metros) en la muestra de tamaño 3.600

Al intentar establecer una comparación entre las formas de las cuatro distribuciones que se tienen, es claro que dicha comparación no se puede realizar adecuadamente si las gráficas están elaboradas en términos de frecuencias absolutas, pues la referencia no es la misma en todos los casos: en el primero, el tamaño de la muestra es 40; en el segundo, es 80; en el tercero, es 200; y, en el cuarto, es 3.600. Es pues, necesario superar el problema de la escala y eso se logra si se hacen las gráficas con base en las frecuencias relativas y no con base en las frecuencias absolutas y además se emplea la misma escala para todas las gráficas. Al tener en cuenta la consideración anterior, se obtiene la siguiente tabla:

Frecuencias relativas (%)				
Estatura	**Muestra 1**	**Muestra 2**	**Muestra 3**	**Población**
1,62	0	0	0,5	0,027
1,64	0	1,25	2	0,5
1,65	7,5	5	2,5	0,47
1,66	0	3,75	2,5	1,38
1,67	0	0	4	4,33
1,68	2,5	6,25	5,5	4,11
1,69	12,5	10	8	5,22
1,70	30	18,75	8	13,47
1,71	0	13,75	9,5	13,55
1,72	25	16,25	10,5	13,63
1,73	0	0	10	13,52
1,74	7,5	7,5	7,5	13,58
1,75	0	0	6	5,05
1,76	12,5	12,5	7,5	4,5
1,77	0	0	5,5	4,11
1,78	2,5	1,25	5	1,36
1,79	0	1,25	3	0,63
1,80	0	2,5	2	0,47
1,82	0	0	0,5	0,027

e. Explique cómo se obtuvo la tabla anterior.

f. Utilice algún criterio para comparar tablas de frecuencias relativas, que le permita establecer en cuál de las muestras, la variable se distribuye de manera más similar a la distribución de la variable en la población.

Con seguridad, usted obtuvo que la tabla más similar a la de la población corresponde a la muestra de mayor tamaño, resultado que es evidente al observar las correspondientes gráficas, si éstas se han hecho con la misma escala y empleando frecuencias relativas.

A continuación se establecen algunos comentarios acerca de las cuatro gráficas anteriores. A medida que se aumenta la cantidad de información, se observa que:

- El perfil de la curva se vuelve cada vez más suave, es decir, se disminuye el número de picos. (En los diagramas realizados no está explícito el perfil de la curva, pero usted lo puede obtener de la siguiente manera: señale el punto medio del lado superior de cada uno de los rectángulos que conforman el diagrama y para cada par de rectángulos consecutivos una sus correspondientes puntos medios.)
- Los valores de la estatura se van agrupando alrededor del valor 1,72 metros, y este valor se ubica en la mitad entre el mínimo y el máximo del conjunto de estaturas.
- La frecuencia de los valores de la estatura disminuye a medida que éstos se alejan de la estatura promedio.
- La curva tiende a ser simétrica con respecto a la vertical que pasa por el valor de la estatura promedio.

Una aproximación al modelo

Se ha visto que a medida que aumenta el tamaño de la muestra, la distribución de la variable "estatura de los estudiantes varones de la universidad" tiende a comportarse de una determinada manera; es esa tendencia la que determinará el modelo que estamos buscando.

Imaginemos que el conjunto de 3.600 estaturas al que nos hemos referido como la población de datos de este caso, en realidad es tan sólo una muestra representativa de una población "muy grande". Veamos cómo se puede esperar que sea la distribución de la estatura en esa población "muy grande". Además, supongamos que se pueden tomar medidas muy finas de la estatura, no hay aproximaciones.

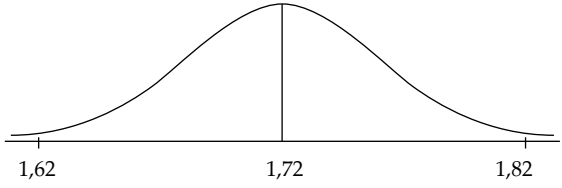
1.- La variable asume todos los valores del intervalo, cuyo extremo inferior es la estatura mínima en la población y cuyo extremo superior es el valor de la estatura máxima en la población. Esto ocurre porque la variable es continua, y en la gráfica se refleja en el hecho de que la curva no tiene "valles".

2.- La estatura promedio está en el centro del intervalo en el que se mueve la variable. Además, es el valor de la variable más frecuente y alrededor del cual se agrupa la mayoría de las observaciones.

3.- La curva es simétrica con respecto a la vertical que pasa por el valor de la estatura promedio: a lado y lado de la estatura promedio, la variable se distribuye de igual manera.

4.- Entre más alejado de la estatura promedio está un valor de la variable, menos frecuente es ese valor en la distribución y recíprocamente, entre menos frecuente sea un valor de la variable, más alejado estará del promedio.

De acuerdo a las características que se espera que tenga la distribución de la variable en la población imaginada, la gráfica de dicha distribución debe ser similar a ésta:



Estatura (metros)
población ideal –modelo–

Y es relativamente fácil aceptar que las otras tres variables mencionadas en este capítulo (promedio ponderado, presión arterial y peso) tienen una distribución similar a la de la estatura, cuando se piensa en una población inmensamente grande.

a. Escoja una de las tres variables (promedio ponderado, presión arterial y peso) y explique el significado de la afirmación:

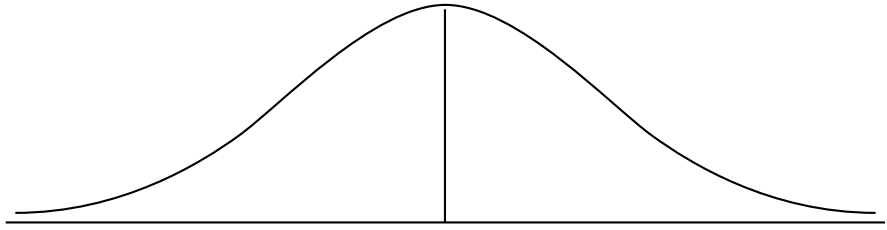
Esa variable se distribuye, en una población muy grande, de manera similar a como lo hace la variable estatura.

Una pausa para resumir

En este punto del proceso hagamos una pausa para resumir lo que se ha realizado. En primer lugar, se formularon dos problemas referidos a una población, problemas que pueden ser complicados y hasta imposibles de resolver, si no se tiene un cierto conocimiento de la estadística, pues en la mayoría de los casos, la información que se requiere para la solución es inaccesible. En segundo lugar, se hizo un análisis del comportamiento de una de las variables involucradas en los problemas planteados, dentro de la correspondiente población y se encontró que tiene una bien determinada distribución. Se concluyó que la distribución de otras variables es similar a la de la estatura. Finalmente se llegó hasta obtener la forma de la curva que representa la distribución de esas variables en poblaciones hipotéticas (inmensamente grandes). Y ahí vamos.

El objetivo es llegar a encontrar un modelo que sea adecuado para representar el comportamiento de variables que se distribuyan de manera similar a como lo hacen las variables con las que hemos trabajado, y además que tal modelo permita responder preguntas que de otra manera es imposible responder.

Parece ser que una curva como la que se presenta a continuación es un buen modelo.



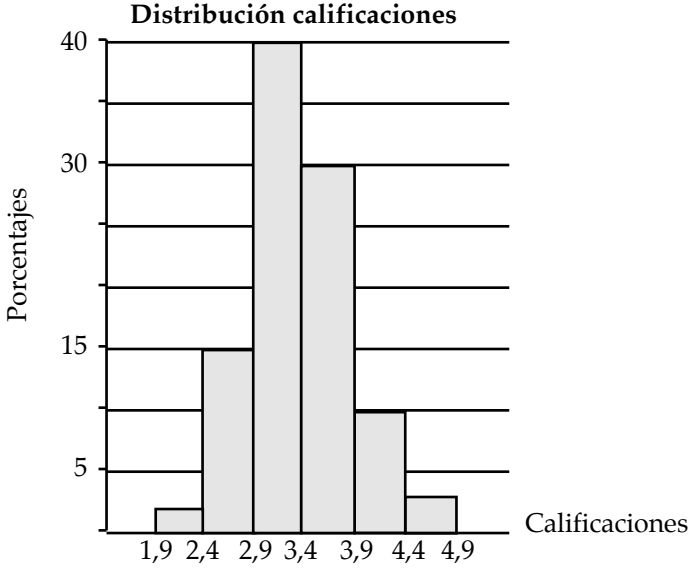
Sin embargo, surgen algunas preguntas:

- ¿De qué manera brinda esa curva información sobre proporciones específicas en la población?
- ¿Cualquier curva "acampanada" o "monticular" es un buen modelo para lo que se desea? Y, si la respuesta es negativa, ¿qué características adicionales deben establecerse sobre tal curva?

Para poder responder las preguntas anteriores se requiere establecer la relación que existe entre probabilidad, proporción y área bajo una curva de distribución de frecuencias. En la siguiente sección vamos a hacer explícita tal relación.

La probabilidad como área bajo una curva

El siguiente histograma representa la distribución de frecuencias agrupadas de las calificaciones obtenidas en un examen de estadística por cien alumnos de la universidad.



a. Determine la proporción de alumnos de la muestra que obtuvieron calificación entre 2,4 y 3,4.

b. Explique cómo obtuvo el resultado a la pregunta anterior.

c. En el histograma sombree los rectángulos para los cuales se cumple que el valor de la variable está entre 2,4 y 3,4. Determine qué proporción del área del histograma es el área sombreada.

d. Calcule la probabilidad de que si se extrae, al azar, un estudiante de esa muestra, su calificación esté entre 2,4 y 3,4.

e. Lea cuidadosamente las preguntas anteriores junto con las respuestas que usted les dio. A partir de eso, escriba una afirmación que relacione, para el caso en cuestión, los siguientes tres conceptos:

- proporción de observaciones
- proporción de área
- probabilidad de que ocurra un determinado evento

A partir de los ejercicios anteriores, usted debió darse cuenta de dos puntos importantes:

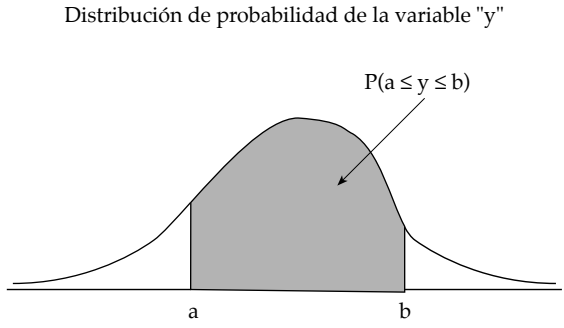
1.- La información que aporta cada rectángulo construido sobre cada clase se puede interpretar de dos maneras:

- como la fracción de observaciones que caen en tal clase
- como la probabilidad de que una observación, extraída al azar de la muestra, caiga en tal clase.

2.- La proporción de observaciones que caen dentro de una clase puede encontrarse en términos de la altura de la correspondiente clase o, también, en términos del área de dicho rectángulo (dado que todos los rectángulos tienen el mismo ancho).

Según lo dicho anteriormente, al tener un histograma que representa la distribución de una variable, es posible calcular la probabilidad de que la variable cumpla una cierta condición, en términos de la correspondiente fracción del área del histograma. Es por eso por lo que, de manera natural, se puede pensar en extender esa idea al caso en que la representación gráfica de la distribución sea una curva continua, sin saltos. Y en realidad así es como se define la distribución de probabilidad de una variable aleatoria continua. En este texto, aceptaremos ese hecho sin entrar en detalles; lo que interesa es saber manejar ese tipo de distribución.

> La probabilidad de que una variable aleatoria continua y, tome algún valor entre a y b, es el área de la región bajo la curva, limitada por los valores a y b, expresada esa área como una fracción o un porcentaje del área total bajo la curva. Lo anterior se nota como $P(a < y < b)$.



La curva normal

En la sección titulada "Una pausa para resumir", quedaron planteadas dos preguntas:

- ¿De qué manera brinda esa curva —la que se ha encontrado como modelo— información sobre proporciones específicas en la población?
- ¿Cualquier curva "acampanada" o "monticular" es un buen modelo para lo que se desea? Y, si la respuesta es negativa, ¿qué características adicionales deben establecerse sobre tal curva?

Pues bien, con la consideración hecha en la sección "La probabilidad como área bajo una curva", se da respuesta a la primera de ellas.

Para considerar que se tiene completamente construido el modelo del que hemos estado hablando, es indispensable dar respuesta a la segunda pregunta: ¿sirve cualquier curva acampanada?

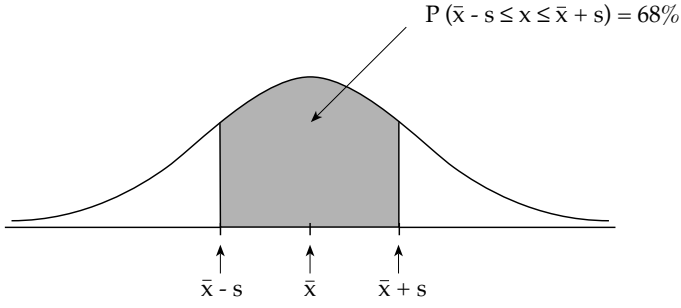
a. ¿Piensa usted que la respuesta a esa pregunta es afirmativa? ¿Por qué?

Al describir el modelo, una de las características se expresó así: "la mayoría de las observaciones se agrupan alrededor del promedio". Sin embargo, cuando

se quiere concretar esa condición se presentan problemas: ¿qué porcentaje del total de observaciones corresponde a la mayoría? ¿Acaso, el 50%? O, ¿el 60%? O, ¿el 80%? Es claro que la respuesta que se dé no es única y en todo caso depende del criterio subjetivo de quien responda. Para concretar la situación debemos decir que vamos a trabajar con una cierta curva, a partir de la cual se determina con alto grado de precisión, el porcentaje del total de observaciones que debe haber entre el valor del promedio y cualquier valor de la variable.

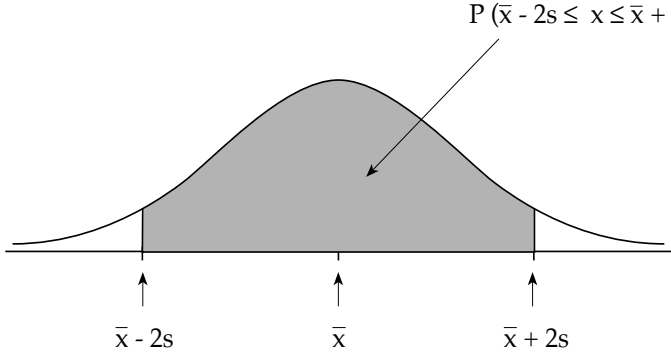
Veamos algunos casos especiales. En el modelo del cual estamos hablando se cumple, por ejemplo, que:

- 68,26% del total de observaciones se encuentra en el intervalo entre 1 desviación estándar antes y 1 desviación después de la media.



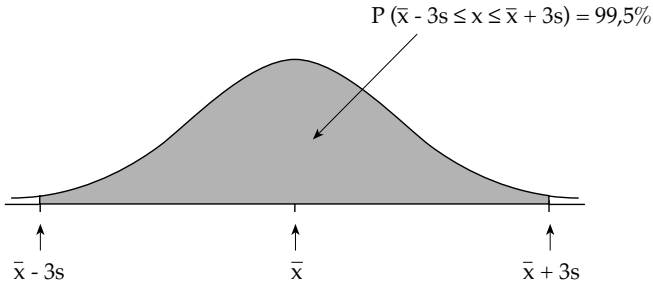
Valores de la variable

- 95,54% del total de observaciones se encuentra en el intervalo entre 2 desviaciones estándar antes y 2 desviaciones después de la media.



Valores de la variable

- 99,74% del total de observaciones se encuentra en el intervalo entre 3 desviaciones estándar antes y 3 desviaciones después de la media.



Un ejemplo puede aclarar lo dicho anteriormente. Se sabe que una variable se distribuye según el modelo del que estamos hablando. Se tomó una muestra de tamaño 50 y se encontró que la media de la variable es 43 y la desviación estándar es 6. Entonces, debe cumplirse que:

- En el intervalo cuyos extremos son: (43 - 6) y (43 + 6) debe haber un 68,26% del número total de observaciones en la muestra. Es decir, 34 de las observaciones de la muestra deben ser valores de la variable que oscilan entre 37 y 49.
- En el intervalo cuyos extremos son: (43 - 12) y (43 + 12) debe haber un 95,54% del número total de observaciones en la muestra. Es decir, 47 ó 48 de las observaciones de la muestra deben ser valores de la variable que oscilan entre 31 y 55.
- En el intervalo cuyos extremos son: (43 - 18) y (43 + 18) debe haber un 99,74% del número total de observaciones en la muestra. Es decir, 49 ó 50 de las observaciones de la muestra deben ser valores de la variable que oscilan entre 25 y 61.

El 68%, el 95% y el 99,5% (valores aproximados) son sólo tres valores especiales asociados con nuestro modelo. Quien conozca el modelo se sabe de memoria dichos valores, pero esos no son los únicos valores asociados al modelo. En realidad, hay una tabla que completa la información con respecto al área de la región bajo la curva comprendida entre el promedio de la distribución y cualquier otro valor de la variable. Tal tabla se presentará más adelante.

Lo que se ha dicho con respecto al área bajo la curva debe conducir a la conclusión de que no toda curva "acampanada" sirve como el modelo buscado.

En resumen, el modelo que vamos a emplear para representar la distribución de ciertas variables continuas, en poblaciones inmensamente grandes tiene las siguientes características:

1.- La variable asume todos los valores reales, es decir, va de $-\infty$ a ∞.

2.- La curva es simétrica con respecto a la vertical que pasa por la media de la variable: a lado y lado de la media, la variable se distribuye de igual manera.

3.- El valor de la media, la moda y la mediana coinciden.

4.- La forma de la distribución de la variable es "acampanada".

5.- El área bajo la curva siempre se distribuye de la misma manera y esto se expresa detalladamente en una tabla.

6.- En la curva, a lado y lado del promedio, hay dos puntos especiales llamados *puntos de inflexión* los cuales marcan un cambio de concavidad en la curva. Esos puntos son los que se asocian a los valores de la variable cuya distancia al promedio es igual a 1 desviación estándar.

La curva que se ha descrito se llama *curva normal* y es posible definirla en términos de una ecuación. Sin embargo, en este texto más que la ecuación, nos interesa conocer las características de la curva y emplearlas cada vez que sea posible para hacer inferencia estadística.

¿Existe una única curva normal?

a. ¿Cree usted que existe una única curva normal? Explique.

b. Si usted sabe que una variable se distribuye normalmente, ¿de qué información requiere para tener una completa imagen gráfica de la distribución?

En una muestra, la variable X —continua— se distribuye normalmente, con media igual a 3,8 y desviación estándar igual a 0,4. En otra muestra, la misma variable se distribuye normalmente con media igual a 4,5 y desviación estándar igual a 0,4.

c. Represente gráficamente, en un mismo plano, la distribución de la variable X en cada una de las dos muestras. Compare las dos curvas.

d. Al mantener igual el valor de la desviación estándar, y variar el valor de la media de la variable, ¿qué efecto se produce en la gráfica? Explique.

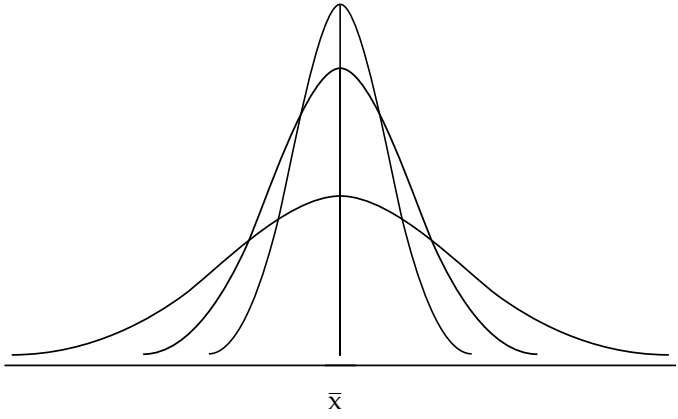
En una muestra, la variable X —continua— se distribuye normalmente, con media igual a 3,8 y desviación estándar igual a 0,8. En otra muestra, la misma variable se distribuye normalmente con media igual a 3,8 y desviación estándar igual a 0,2. En una tercera muestra, la misma variable se distribuye normalmente con media igual a 3,8 y desviación estándar igual a 1,4.

e. Represente gráficamente, en un mismo plano, la distribución de la variable X en cada una de las tres muestras. Compare las tres curvas.

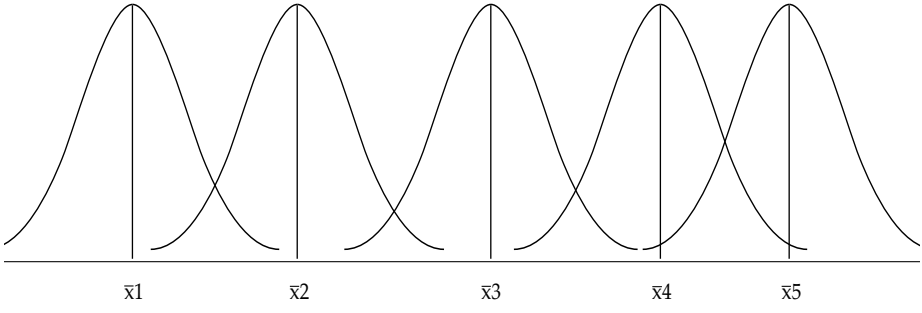
f. Al mantener igual el valor de la media y variar el valor de la desviación estándar de la variable, ¿qué efecto se produce en la gráfica? Explique.

La curva normal depende de dos valores: la media y la desviación estándar de la variable. Esto quiere decir que si una variable se distribuye normalmente y si la media es $\bar{x}$, dependiendo del valor de la desviación estándar de la variable, se obtendrán tantas curvas normales como valores haya para la desviación estándar. Esas curvas serán más o menos altas y por tanto menos o más anchas según el valor de la desviación estándar. En total, si se dieran todos los valores posibles a la desviación estándar y se mantuviera fijo el valor de la media, se obtendría una familia de curvas normales, las cuales diferirían entre s, sólo en su altura y su anchura. Pero la distribución del área bajo cualquiera de ellas **siempre** sería la misma. También es cierto que si una variable se distribuye normalmente y si la desviación estándar es s, dependiendo del valor de la media de la variable, se obtendrán tantas curvas normales como valores haya para la media. En total, si se dieran todos los valores posibles a la media y se mantuviera fijo el valor de la desviación estándar, se obtendría una familia de curvas normales, las cuales diferirían entre sí, sólo en la ubicación sobre el eje horizontal; de resto todas serían idénticas en forma. Y la distribución del área bajo cualquiera de ellas **siempre** sería la misma.

Lo dicho en los dos párrafos anteriores, se puede esquematizar de la siguiente manera, y permite concluir que **no** existe una única curva normal.



Tres distribuciones normales: media $\bar{x}$ y diferente desviación estándar



Cinco distribuciones normales: desviación estándar fija y diferentes medias

Y... el modelo

La región bajo la curva normal puede ser alta y delgada o corta y ancha, según el valor de la desviación estándar y también según la relación que exista entre las escalas horizontal y vertical empleadas al hacer la gráfica. Pero como ya se ha dicho antes, la distribución del área bajo la curva **siempre** es la misma. Es esa la razón por la cual, para determinar el área bajo la curva, comprendida entre dos valores cualesquiera de la variable, no hace falta tener tantas tablas como diferentes curvas normales. Es suficiente disponer de una única tabla que exprese las áreas de una curva normal, a la cual pueda reducirse cualquier otra curva normal, sin que importe en cada caso el valor particular del promedio ni el de la desviación estándar.

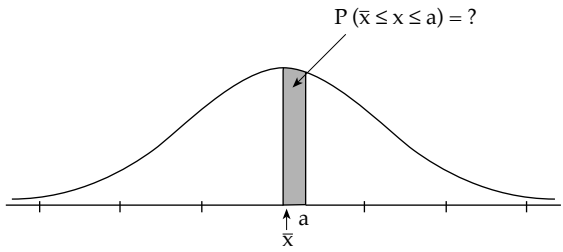
El problema, es entonces, decidir cuál es la curva normal más adecuada para adoptarla como la curva normal patrón o, dicho de otra manera, como la *curva normal estándar*. Y, la decisión estará tomada cuando se encuentren los valores más apropiados para la media y la desviación estándar.

Las preguntas que se formulan a continuación tienen por objeto promover la reflexión sobre cuáles son los valores más convenientes que deben asignarse a la media y a la desviación estándar de la distribución normal estándar.

a. Con respecto a la siguiente afirmación, determine si usted está de acuerdo con ella. Explique su respuesta.

> Al hacer la tabla que describa la distribución del área, no es indispensable que se describa la distribución del área bajo **toda** la curva.

Es suficiente describir la distribución del área para la porción de la curva situada, por ejemplo, a la derecha del promedio, pues como la curva es simétrica con respecto a la vertical que pasa por el valor del promedio, las regiones situadas a lado y lado son idénticas. Entonces la información que debe aportar la tabla es el porcentaje de área bajo la curva, entre el promedio y un valor de la variable, (mayor que el promedio, o menor que el mismo). Aceptemos que la tabla indica cuál es el área (relativa) bajo la curva entre el valor del promedio y cualquier valor de la variable que sea mayor que el promedio.



b. ¿Se podría asignar, por ejemplo, el valor 4 al promedio? Y, ¿el valor 10? Y, ¿el valor 0? De los anteriores valores, ¿cuál es el más adecuado? Explique.

c. ¿En cuál de los tres casos mencionados en la pregunta anterior, resulta más fácil medir la distancia del promedio a un determinado valor de la variable?

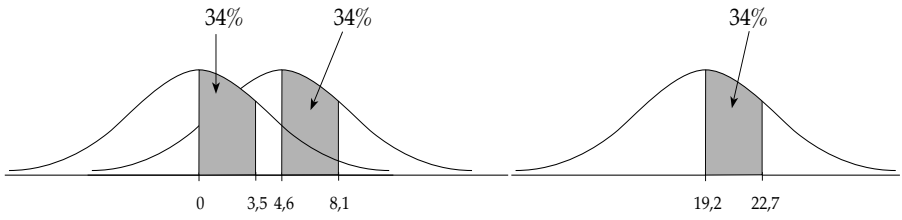
El valor de la media más adecuado para definir la distribución normal estándar es 0, pues al tener que medir distancias de un valor a otro tomado como referencia, el caso más sencillo se da cuando la referencia es 0. Veámoslo en un ejemplo. Suponga que se tienen tres variables distribuidas normalmente, que la desviación estándar de las tres variables es la misma, por ejemplo, 3,5 y que tienen diferente media. Para cada una de las tres distribuciones, se quiere determinar el área bajo la curva, comprendida entre el correspondiente promedio y el valor dado a. En la tabla siguiente se da la media de cada una de las tres variables y también se da el valor de a.

Distribución	**Media**	**Valor de** a
1	4,6	8,1
2	19,2	22,7
3	0	3,5

Se tiene pues que en los tres casos, la distancia del promedio al valor de *a* es la misma, 3,5, porque:

- 8,1 - 4,6 = 3,5
- 22,7 - 19,2 = 3,5
- 3,5 - 0 = 3,5

y por tanto el área bajo la curva es, en los tres casos, la misma: 34%. La explicación es la siguiente: se sabe que un 68% del área total corresponde al área de la región bajo la curva, que está limitada por los valores que distan 1 desviación estándar del valor del promedio. Por tanto, como lo que nos interesa es la mitad de tal región, su área será la mitad de 68%.



Es importante observar que aunque la respuesta es sencilla de obtener en los tres casos, la situación más sencilla, más natural se tiene en el caso de la distribución 3. En los tres casos, se hace necesario **calcular la distancia** del *valor específico* de la variable al valor de la media de la variable. Para hacer el cálculo se hace una resta. Sin embargo, la situación en la que la resta es inmediata es en el caso en el que la referencia con respecto a la cual se está calculando la distancia es 0. Por esa razón, para construir la distribución normal estándar preferimos que el valor de la media sea 0 y no otro valor.

Para definir completamente la distribución normal estándar hace falta determinar cuál es el valor de la desviación estándar más adecuado. Encontrémoslo.

d. Explique por qué no se podría asignar el valor 0 a la desviación estándar, para obtener una distribución normal.

e. ¿Se podría asignar el valor 4 a la desviación estándar? Y, ¿el valor 10? Y, ¿el valor 1? De los anteriores valores, ¿cuál es el más adecuado? Explique.

f. ¿En cuál de los tres casos mencionados en la pregunta anterior, se tiene la unidad de medida más cómoda?

El valor de la desviación estándar más adecuado para definir la distribución normal estándar es 1. Veámoslo en un ejemplo. Suponga que se tienen tres variables distribuidas normalmente, que la media de las tres variables es la misma, por ejemplo, 0, y que tienen diferente desviación estándar. Para cada una de las tres distribuciones, se quiere determinar el área bajo la curva, com-

prendida entre el promedio y el valor dado a. En la tabla siguiente se da la desviación estándar de cada una de las tres variables y también se da el valor de a.

Distribución	Desviación	Valor de a
1	2,8	5,6
2	0,8	1,6
3	1	2

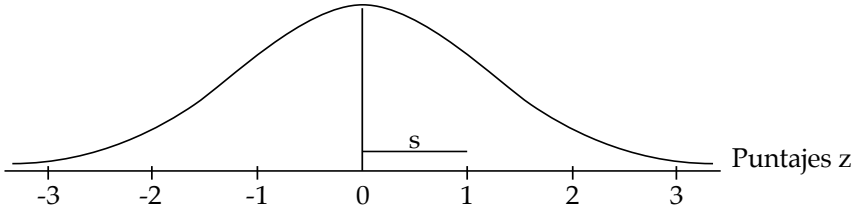
Se tiene pues que en los tres casos, la distancia del promedio al valor de a, **medida en términos de la correspondiente desviación estándar** es la misma, 2, porque:

- (5,6 - 0) / 2,8 = 2
- (1,6 - 0) / 0,8 = 2
- (2 - 0) / 1 = 2

y por tanto el área bajo la curva es, en los tres casos, la misma: 47,72%. La explicación es la siguiente: se sabe que un 95,44% del área total corresponde al área de la región bajo la curva, que está limitada por valores que distan 1,96 desviaciones estándar del valor del promedio; por tanto, como lo que nos interesa es la mitad de tal región, su área será la mitad de 95,44%.

Es importante observar que aunque la respuesta es sencilla de obtener en los tres casos, la situación más sencilla, más natural se tiene en el caso de la distribución 3. En los tres casos se hace necesario **calcular la distancia** del *valor específico* de la variable al valor de la media de la variable y **medir dicha distancia en términos de la correspondiente desviación estándar**. Para hacer el cálculo se hace una división. Sin embargo, la situación en la que la división es inmediata es en el caso en el que la unidad de medida, o sea la desviación estándar es 1. Por esa razón, para construir la distribución normal estándar preferimos que el valor de la desviación estándar sea 1 y no otro valor.

En efecto, si la distribución es normal con media igual a 0 y desviación estándar igual a 1, se habla de la *distribución normal estándar*. Esta distribución representa una idealización del comportamiento de cualquier variable que se distribuya normalmente.



Distribución normal estándar

Con respecto a la última gráfica, los valores que aparecen en el eje horizontal no son valores de alguna variable; se llaman *puntajes estandarizados* o *puntajes z* y siempre se puede establecer una correspondencia **biunívoca** entre los valores de cualquier variable que se distribuya normalmente y dichos puntajes. Es decir, siempre que se tenga una variable distribuida normalmente cada valor de esa variable tiene asociado un único valor de la distribución normal estándar y todo valor de la distribución normal estándar tiene asociado un único valor de la variable. A cada valor negativo de la distribución normal estándar corresponde un bien determinado valor de la variable, inferior al valor del promedio, mientras que a cada valor positivo de la distribución normal estándar corresponde un cierto valor de la variable, superior al valor del promedio.

Manejo de la distribución normal estándar

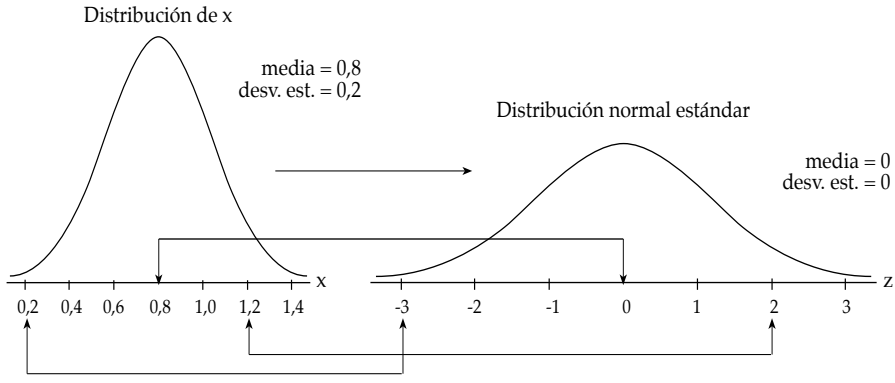
¿Cómo se realiza esa correspondencia biunívoca entre valores de la variable y los puntajes z? Existen dos procesos, inversos entre sí, mediante los cuales se establece la correspondencia de la cual hemos venido hablando. Vamos a llamarlos estandarización y desestandarización. A continuación se aclarará el significado y la forma de operar con dichos procesos.

Estandarización

Una variable X se distribuye normalmente con media igual a 0,8 y desviación estándar igual a 0,2. La tabla que se presenta a continuación muestra algunas correspondencias entre valores de la variable X y puntajes z.

valor de la variable	0,2	0,4	0,6	0,8	1	1,2	1,4
puntajes z	-3	-2	-1	0	1	2	3

Además, en la siguiente gráfica se muestran dichas correspondencias.



Correspondencia entre algunos valores de la variable
y los respectivos puntajes z

a. Intente explicar cómo se obtienen esas correspondencias.

Puede parecer que en algunos casos es fácil hallar la correspondencia, mientras que en otros casos no. (Por ejemplo, se puede encontrar, por simple inspección, el puntaje z asociado al valor 1,4 de la variable, pues 1,4 y la media de la distribución (0,8) están a una distancia de (0,6) lo que corresponde a una distancia de 3 veces la correspondiente desviación estándar. Como 1,4 es superior a la media de la distribución entonces el puntaje z asociado con 1,4 es (+3). Sin embargo, no es tan evidente cuál es el puntaje z asociado al valor 1,16 de la variable.) Entonces debemos encontrar una forma que permita determinar la correspondencia entre un valor de la variable y su respectivo puntaje z, en cualquier caso.

Para la distribución de la variable X definida anteriormente, usted va a encontrar el puntaje z asociado a 0,43. Para ello, responda las siguientes preguntas:

b. Haga la gráfica de la distribución de la variable X, y en ella localice el promedio y el valor 0,43. Además, dibuje la distribución normal estándar y en ella localice la media (0) y ubique de manera aproximada el puntaje z asociado con 0,43.

c. ¿El puntaje z, asociado a 0,43, es positivo o negativo? ¿Por qué?

d. En la distribución de la variable, ¿qué distancia hay de 0,43 al promedio? ¿Cómo calcula esa distancia? En la operación que acaba de realizar, ¿cómo puede indicarse el hecho de que el valor 0,43 es menor que el promedio de la variable? (Tenga en cuenta ese hecho para explicar cómo calcula la "distancia" entre el promedio y el valor específico de la variable.)

e. Comente la siguiente afirmación:

> La distancia que hay entre 0,43 y 0,8 debe ser la misma que haya entre 0 y el puntaje z que se busca.

f. En la distribución de la variable X, ¿cuál es la unidad de medida? Y, en la distribución normal estándar, ¿cuál es la unidad de medida?

g. Con base en la respuesta anterior, revise su comentario referente a la comparación de distancias.

En realidad, sí interesa saber cuál es la distancia que hay entre el promedio y el valor 0,43, pero además de eso interesa saber cuánto es esa distancia en términos de la unidad de medida, que es la desviación estándar y que para el caso es 0,2. Dicho de otra manera, es necesario averiguar *cuántas veces cabe*, la desviación estándar de la distribución de la variable, en esa distancia.

h. Calcule la "distancia" en unidades de desviación estándar que hay entre el valor 0,43 de la variable y el promedio de la misma.

i. ¿Qué distancia, medida en desviaciones estándar de la correspondiente distribución, debe haber entre la media (0) y el puntaje z? Entonces, ¿qué valor asume ese puntaje z?

Puesto que las distancias, medidas en unidades de la correspondiente desviación estándar, entre la media y el valor específico de la variable (en la distribución de la variable) y la media y el puntaje z (en la distribución normal

estándar) deben ser iguales y además la desviación estándar de la segunda distribución mencionada es 1, entonces se deduce que la distancia entre la media de esa distribución (0) y el puntaje z debe ser el número que se encontró en el item anterior.

j. Explique tan claramente como le sea posible el significado de la siguiente frase:

El puntaje z asociado al valor 1,16 de la variable X es 1,8.

k. Compruebe que el puntaje z asociado al valor 1,16 de la variable X es 1,8. (Tenga en cuenta su respuesta al item anterior.)

l. Explique tan claramente como le sea posible el significado de la siguiente frase:

El puntaje z asociado al valor 0,54 de la variable X es -1,3.

m. Compruebe que el puntaje z asociado al valor 0,54 de la variable X es -1,3. (Tenga en cuenta su respuesta al item anterior.)

n. Escriba una receta que le permita encontrar el puntaje z asociado a cualquier valor de la variable.

o. Encuentre el valor de la variable, que es superior al promedio de la misma y cuya distancia a él sea de 2,67 veces el valor de la desviación estándar.

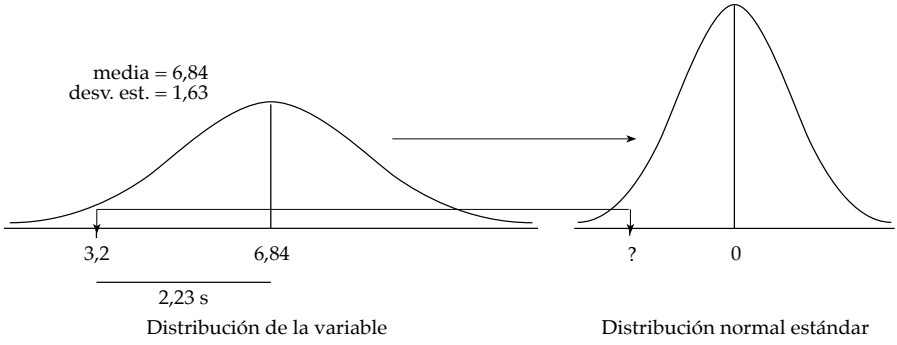
p. Encuentre el valor de la variable, que es inferior al promedio de la misma y cuya distancia a él sea de 1,38 veces el valor de la desviación estándar.

Ahora vamos a desarrollar completamente un ejemplo donde se puedan aclarar dudas con respecto al proceso que permite pasar de la distribución de una variable normal a la distribución normal estándar.

Suponga que una variable se distribuye normalmente, tiene media igual a 6,84 y desviación estándar igual a 1,63. Se pide encontrar los puntajes z asociados a los valores 3,2 y 7,9 de la variable.

Encontremos el puntaje z asociado al valor 3,2. Para ello vamos a realizar los siguientes pasos:

- Representamos gráficamente la situación: es decir, en una curva normal localizamos la media de la distribución de la variable y el valor específico de la variable (3,2). En otra curva normal, representamos la distribución normal estándar y en ella localizamos la media y el puntaje z que estamos buscando.



- En la distribución de la variable, calculamos la distancia que hay entre 3,2 y el valor de la media. Para ello, hacemos la diferencia entre ese par de valores. Sin embargo, aquí es importante indicar si el valor específico de la variable (3,2) es menor o mayor que el valor de la media. Para lograrlo, interesa que el resultado de la diferencia tenga la posibilidad de ser mayor o menor que cero, según cuál sea el caso. Es decir: si el valor de la variable es menor que el promedio, la "distancia" debería resultar negativa; y si el valor de la variable es mayor que el promedio, la distancia debería resultar positiva. Por eso, para calcular la "distancia" entre el valor específico de la variable y la media se realiza la diferencia

valor de la variable - valor del promedio

En el caso que nos ocupa, esa "distancia" es: 3,2 - 6,84 = -3,64

- Expresamos la "distancia" que hay del valor específico de la variable a la media en unidades de desviación estándar, es decir, calculamos

cuántas veces "cabe" la desviación estándar en esa "distancia". Para ello se divide el valor de la "distancia" entre el valor de la desviación estándar

(valor de la variable - valor del promedio) / (desv. est.)

En el caso que nos ocupa, se obtiene: -3,64 / 1,63 = -2,23

- En la distribución normal estándar, el puntaje z que estamos buscando dista de la correspondiente media (0), tanto como, en la distribución de la variable, el valor específico dista de la media. Por tanto, como la media de la distribución de puntajes z es 0, el puntaje z buscado es precisamente el último valor encontrado

(valor de la variable - valor del promedio) / (desv. est.)

En el caso que nos ocupa, el puntaje z asociado a 3,2 es -2,23. Y, se interpreta así: el valor de la variable (3,2) está a la izquierda, y a 2,23 unidades de desviación estándar del valor del promedio.

Ahora bien, encontremos el puntaje z asociado al valor 7,9 de la variable.

- Primero, representamos gráficamente la situación.
- Segundo, calculamos la distancia que hay de 7,9 a 6,84:

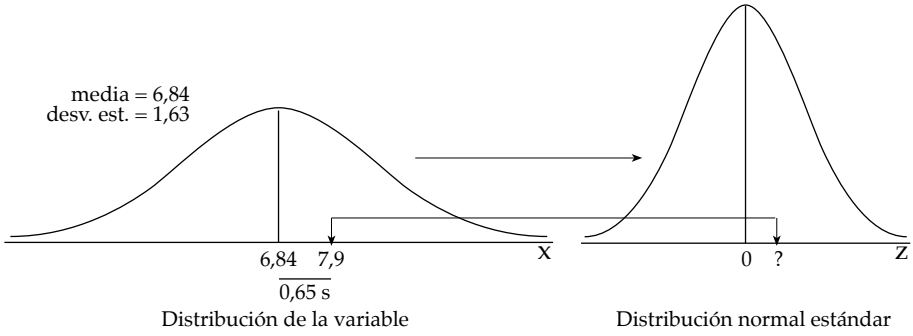
7,9 - 6,84 = 1,06

- Expresamos esa distancia en unidades de desviación estándar:

1,06 / 1,63 = 0,65

- Identificamos el valor 0,65 con el puntaje z que buscamos. Y, se interpreta así: el valor 7,9 de la variable está a la derecha y a 0,65 unidades de desviación estándar del valor del promedio.

El esquema de la situación se muestra a continuación:



Formalicemos un poco lo que se ha hecho en esta sección.

> La correspondencia biunívoca que existe entre los valores de una distribución normal cualquiera y los puntajes z de la distribución normal estándar se establece a través de un proceso llamado *estandarización*, el cual se puede resumir mediante la siguiente expresión:
> (Valor de la variable - media) / desviación estándar = puntaje z

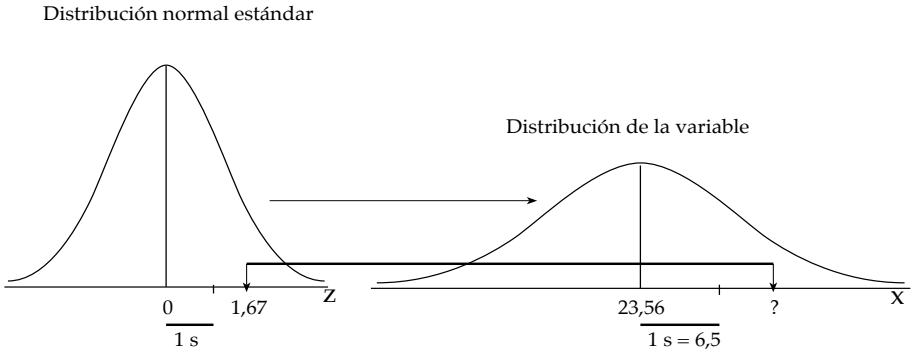
La reversa en el proceso de estandarización

De acuerdo al resultado obtenido anteriormente, a cada valor de una variable —distribuida normalmente— le corresponde un único puntaje z de la distribución normal estándar, mediante la expresión (valor de la variable - media) / desviación estándar. Ahora, encontremos (no de manera mecánica, sino dando significado al proceso) la expresión que permite asociar a cada puntaje z de la distribución normal estándar un bien determinado valor de cualquier variable distribuida normalmente.

Una variable se distribuye normalmente, con media igual a 23,56 y desviación estándar igual a 6,5. Se sabe que el puntaje z asociado a un cierto valor de la variable es 1,67. Se quiere encontrar el valor de la variable. Para ello vamos a realizar los siguientes pasos:

- Representamos gráficamente la situación: es decir, en la curva normal estándar localizamos la media de la distribución (0) y el puntaje z = 1,67. Dicho puntaje se ubica a la derecha de la media, por tanto, en la

distribución de la variable, el valor buscado se encuentra también a la derecha de la correspondiente media (23,56). Teniendo en cuenta esto, en otra curva normal, representamos la distribución de la variable y en ella localizamos la media y señalamos, aproximadamente, el valor que buscamos.

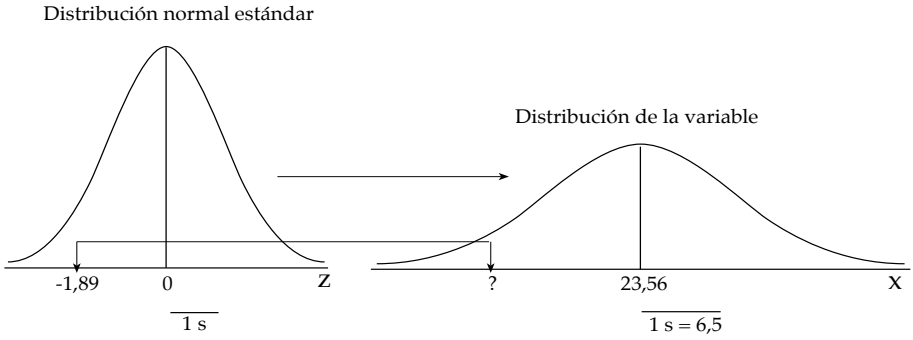


- En la distribución normal estándar, el puntaje z (1,67) dista de la media (0) tanto como, en la distribución de la variable, el valor que buscamos dista de la correspondiente media (23,56). Por tanto, se concluye que la distancia que debe haber entre el valor buscado de la variable y la correspondiente media (23,56) —medida en unidades de desviación estándar— es 1,67 veces la desviación estándar.
- Puesto que la desviación estándar de la distribución de la variable es igual a 6,5, entonces, la distancia entre la media de dicha distribución y el valor buscado es igual a 1,67 x 6,5 = 10,855.
- Si la distancia de la media al valor buscado es de 10,855 y además el valor buscado es mayor que la media, entonces para encontrar dicho valor debemos sumar el valor de la media con el de la distancia. Es decir, 23,56 + 10,855 = 34,415. 34,415 es el valor de la variable que es superior a la media y dista 1,67 de desviación estándar de la media.

Veamos otro ejemplo. Una variable se distribuye normalmente, con media igual a 23,56 y desviación estándar igual a 6,5. Se sabe que el puntaje z asociado a un cierto valor de la variable es -1,89. Se quiere encontrar el valor de la variable.

- Representamos gráficamente la situación: es decir, en la curva normal estándar localizamos la media de la distribución (0) y el puntaje z = -1,89. Dicho puntaje se ubica a la izquierda de la media, por tanto,

en la distribución de la variable, el valor buscado se encuentra también a la izquierda de la correspondiente media (23,56). Teniendo en cuenta esto, en otra curva normal, representamos la distribución de la variable y en ella localizamos la media y señalamos, aproximadamente, el valor que buscamos.



- En la distribución normal estándar, el puntaje z (-1,89) dista de la media (0) tanto como, en la distribución de la variable, el valor que buscamos dista de la correspondiente media (23,56). Por tanto, se concluye que la distancia que debe haber entre el valor buscado de la variable y la correspondiente media (23,56) —medida en unidades de desviación estándar— es 1,89 veces la desviación estándar.
- Puesto que la desviación estándar de la distribución de la variable es igual a 6,5, entonces, la distancia entre la media de dicha distribución y el valor buscado es igual a 1,89 x 6,5 = 12,285.
- Si la distancia de la media al valor buscado es de 12,285 y además el valor buscado es menor que la media, entonces para encontrar dicho valor debemos restar al valor de la media el de la distancia. Es decir, 23,56 - 12,285 = 11,275. 11,275 es el valor de la variable que es inferior a la media y dista 1,89 de desviación estándar de la media.

Formalicemos un poco lo que se ha hecho en esta sección.

Cada puntaje z de la distribución normal estándar se puede asociar con un único valor de una variable distribuida normalmente, mediante un proceso llamado *desestandarización*, el cual se puede resumir mediante la siguiente expresión: media ± (# de desviaciones estándar que hay de la media al valor de la variable) = (valor de la variable)
Es decir, media ± (puntaje z) x (desv. est.) = (valor de la variable)

Tabla de la distribución normal

Las calificaciones de un examen realizado a los estudiantes del curso de estadística se distribuyen normalmente, con media igual a 3,2 y desviación estándar igual a 0,4. Se quiere determinar qué porcentaje del número de alumnos obtuvo calificación entre 2,4 y 4,0. Para ello, responda las siguientes preguntas.

a. Represente gráficamente la distribución de la variable calificación y localice en ella la media y los valores de la variable que se conocen.

b. Estandarice los valores 2,4 y 4,0 y represente esa información en la curva normal estándar.

Seguramente usted encontró que los puntajes z asociados a 2,4 y 4,0 son respectivamente -2 y 2.

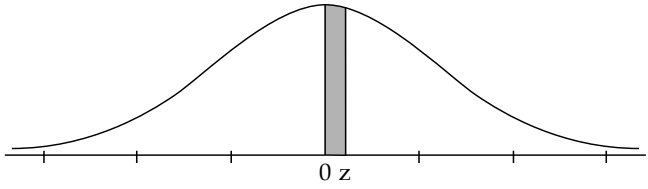
La pregunta "¿qué porcentaje del número de alumnos obtuvo calificación entre 2,4 y 4,0?" puede entonces reformularse así: ¿qué porcentaje del número de alumnos obtuvo calificación entre -2 y 2? La afirmación anterior es cierta gracias a que la distribución del área bajo cualquier curva normal siempre es la misma. Y, así reformulada la pregunta su respuesta es inmediata puesto que sabemos que 95,44% del total de las observaciones se encuentra en el intervalo entre 2 desviaciones antes y 2 desviaciones después de la media.[21] Por tanto,

21 Véase la sección titulada "La curva normal".

95,44% de los alumnos que presentaron el examen obtuvieron calificación entre 2,4 y 4,0.

Si la pregunta hubiera sido, por ejemplo, "¿qué porcentaje del número de alumnos obtuvo calificación entre 2,6 y 3,8?", a pesar de que el proceso de solución es idéntico al anterior, con la teoría que se ha desarrollado hasta ahora no se habría podido responder tal pregunta puesto que los puntajes z asociados a 2,6 y 3,8 respectivamente son -1,5 y 1,5 y no conocemos qué porcentaje del área total corresponde al área bajo la curva normal estándar, entre los valores -1,5 y 1,5.

De lo anterior, debe ser evidente la necesidad de conocer la distribución completa del área bajo la curva normal. Existe una tabla que da esa distribución; ahora el problema se centra en aprender a manejarla. La tabla que emplearemos se presenta al final de este capítulo. Veamos una porción de ella:



z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	,0000	,0040	0,008	,0120	,0160	,0239	,0239	,0279	,0319	,0359
0,1	,0398	,0438	,0478	,0517	,0557	,0596	,0636	,0675	,0714	,0753
0,2	,0793	,0832	,0871	,0910	,0948	,0987	,1026	,1064	,1103	,1141
0,3	,1179	,1217	,1255	,1293	,1331	,1368	,1406	,1443	,1480	,1517
...										
1,4	,4192	,4207	,4222	,4236	,4251	,4265	,4279	,4292	,4306	,4319
1,5	,4332	,4345	,4357	,4370	,4382	,4394	,4406	,4418	,4429	,4441

En primer lugar, la tabla mencionada da el área de la región que queda bajo la curva y comprendida entre el valor de la media (0) y un determinado puntaje z, a la derecha de la media. Esto que se acaba de decir, aparece representado gráficamente en la parte superior de la tabla.

La tabla es un rectángulo de 32 filas por 11 columnas. La primera columna de la tabla (encabezamiento) contiene valores que puede asumir el puntaje z.

Esos valores, que varían de décima en décima, van desde 0.0 hasta 3.0, es decir, algunos de los valores que se dan en esa columna son: 0,0, 0,1, 0,2, 0,3, 0,4, etc. De esa manera se tienen 31 diferentes puntajes z. La segunda columna (encabezada con 0.00) da el área de la región que está bajo la curva y está comprendida entre la media de la distribución (0) y el valor dado del puntaje z. Por ejemplo, mirando esa segunda columna se sabe que el área de la región bajo la curva, comprendida entre 0 y z = 1,60 es 0,4452; también se sabe que el área de la región bajo la curva, comprendida entre 0 y 2,50 es 0,4938; etc.

A pesar de que en la primera columna sólo hay 31 valores para el puntaje z, es posible considerar más de 31 valores para el puntaje z, y por tanto, es posible saber el área para muchos más casos. ¿Cómo? Para eso consideremos la primera fila, la encabezada por z. En esa fila aparecen los valores 0,00, 0,01, 0,02, hasta 0,09. Pues bien, teniendo en cuenta la primera columna y la primera fila de la tabla es posible "armar", por ejemplo, el valor 1,56 para z. Para hacerlo, mire en la primera columna el valor 1,5 y en la primera fila, el valor 0.06; en la intersección de la fila encabezada por 1,5 con la columna encabezada por 0,06, encuentra el área de la región bajo la curva comprendida entre 0 y el puntaje z = 1,56. Esa área es 0,4406.

a. Compruebe que el área de la región bajo la curva, comprendida entre 0 y el puntaje 2,18 es 0,4854.

b. Compruebe que el área de la región bajo la curva, comprendida entre 0 y el puntaje 1,07 es 0,3577.

c. Volvamos a la situación planteada al inicio de esta sección: la calificación de los alumnos en un examen se distribuye normalmente con media igual a 3,2 y desviación estándar igual a 0,4. Se quiere saber qué porcentaje del número de alumnos obtuvo calificación entre 2,6 y 3,8. Al reformular la pregunta, en términos de puntajes z (eso ya se hizo anteriormente) lo que interesa conocer es el área bajo la curva normal comprendida entre los puntajes -1,5 y 1,5.

Verifique que la afirmación siguiente es correcta y explique.

> Un 86,64% del número total de alumnos que presentaron el examen obtuvo calificación entre 2,6 y 3,8.

Notación

Hasta el momento no hemos utilizado ninguna forma de abreviar y de notar el área de la región bajo la curva, comprendida entre 0 y un determinado puntaje z, a la derecha de la media (0). Vamos, entonces, a adoptar dicha convención. Primero, utilizaremos la convención para un par de casos particulares y luego, la emplearemos para el caso general:

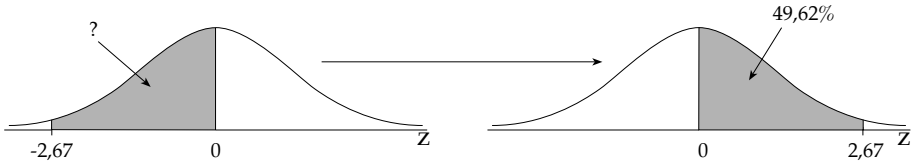
Notación	Significado
$P(0 < z < 1{,}60) = 0{,}4452$	el área de la región bajo la curva, comprendida entre 0 y 1,6 es 44,52%
$P(0 < z < 2{,}50) = 0{,}4938$	el área de la región bajo la curva, comprendida entre 0 y 2,5 es 49,38%
$P(0 < z < a) = A$	el área de la región bajo la curva, comprendida entre 0 y a es A
$P(c < < z < d) = B$	el área de la región bajo la curva, comprendida entre c y d es B

Manejo de la tabla

A pesar de que la tabla sólo da información del área de una región bajo la curva, que está comprendida entre 0 y un puntaje z —positivo—, es posible calcular el área de cualquier región bajo la curva. Para ello se utiliza el hecho de que la curva es simétrica con respecto a la vertical que pasa por 0 y además, la posibilidad de expresar el área de una región como suma de las áreas de varias regiones, o como la diferencia de las áreas de dos regiones. Veamos los siguientes ejemplos.

Ejemplo 1. Determinar el área de la región bajo la curva, que está comprendida entre 0 y el puntaje z = -2,67.

Puesto que la curva normal estándar es simétrica con respecto a la vertical que pasa por 0, la región bajo la curva comprendida entre -2,67 y 0 tiene la misma área que la región bajo la curva comprendida entre 0 y 2,67. Por tanto, como el área de la última región mencionada es 49,62% del área total bajo la curva, entonces el área que estamos buscando es también 49,62% del área total.



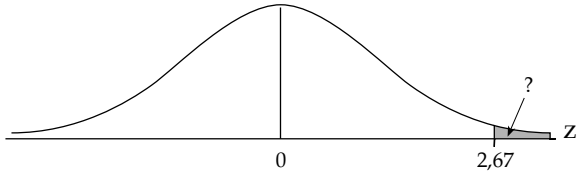
Lo anterior puede notarse de la siguiente manera:

$$P(-2{,}67 < z < 0) = P(0 < z < 2{,}67)$$

$$P(0 < z < 2{,}67) = 0{,}4962$$

Por tanto, $P(-2{,}67 < z < 0) = 49{,}62\%$.

Ejemplo 2. Determinar el área de la región bajo la curva, a la derecha del puntaje z = 2,67.



Para este caso, la tabla no da el área buscada. Por tanto, debemos encontrarla de manera indirecta. Conviene, entonces, expresar el área pedida como suma o diferencia de áreas que se puedan conocer por medio de la tabla. El área de la región que nos interesa se puede expresar como la diferencia de dos áreas.

Puesto que la curva normal estándar es simétrica con respecto a la vertical que pasa por 0, dicha vertical divide la región bajo la curva en dos regiones de igual área y el área de cada una de ellas es el 50% del área total. Por otro lado, se sabe que el área de la región bajo la curva, comprendida entre 0 y el puntaje 2,67 es 49,62%. Si al área de la región bajo la curva, que está a la derecha de la media se le quita el área de la región bajo la curva, comprendida entre 0 y 2,67 se obtiene el área buscada. Es decir, el área buscada es 0,5 - 0,4962 = 0,0038.

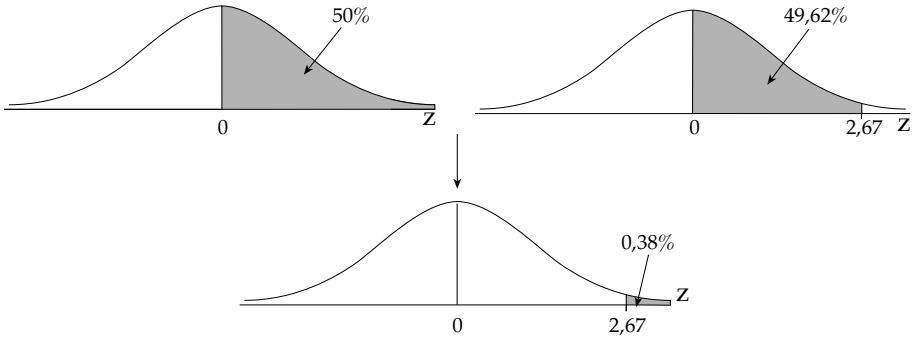
Lo anterior puede notarse de la siguiente manera:

$$P(z > 0) = 0,5$$

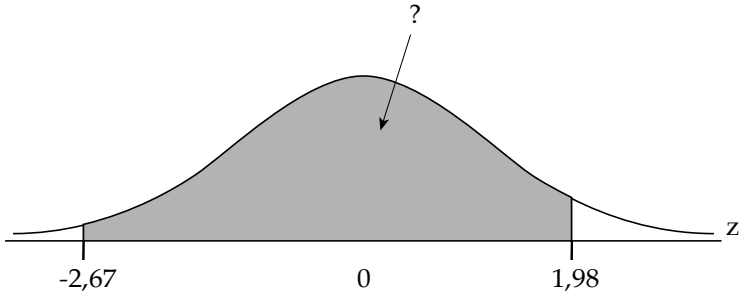
$$P(0 < z < 2,67) = 0,4962$$

$$P(z > 0) - P(0 < z < 2,67) = P(z > 2,67)$$

Por tanto, $P(z > 2,67) = 50\% - 49,62\% = 0,38\%$



Ejemplo 3. Determinar el área de la región bajo la curva, que está comprendida entre los puntajes -2,67 y 1,98.



Nuevamente en este caso, la tabla no da el área buscada. Por tanto, debemos encontrarla de manera indirecta. Conviene, entonces, expresar el área pedida como suma o diferencia de áreas que se puedan conocer por medio de la tabla. El área de la región que aquí nos interesa se puede expresar como la suma de dos áreas.

Podemos descomponer la región cuya área se busca en dos regiones: una, la comprendida entre -2,67 y 0, y la otra región, la comprendida entre 0 y 1,98. De esa manera, el área buscada será la suma de las áreas de las dos regiones antes mencionadas. Por el ejemplo 1, sabemos que el área de la región comprendida entre -2,67 y 0, es 49,62% del área total. Y buscando directamente en la tabla se encuentra que el área de la región comprendida entre 0 y 1,98 es 47,61% del área total. Por tanto, el área buscada es la suma de 49,62% y 47,61%.

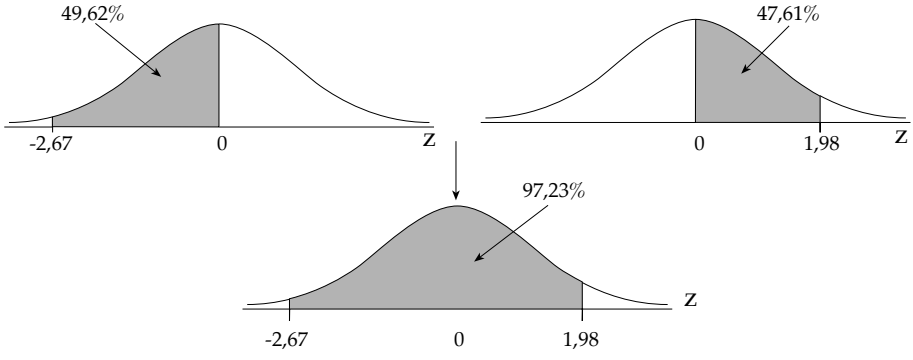
Lo anterior puede notarse de la siguiente manera:

$$P(-2{,}67 < z < 1{,}98) = P(-2{,}67 < z < 0) + P(0 < z < 1{,}98)$$

$$P(-2{,}67 < z < 0) = P(0 < z < 2{,}67) = 49{,}62\%$$

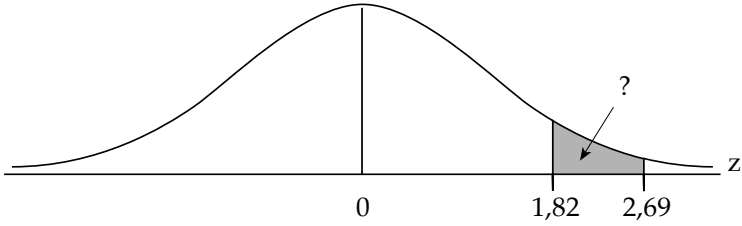
$$P(0 < z < 1{,}98) = 47{,}61\%$$

Por tanto, $P(-2{,}67 < z < 1{,}98) = 49{,}62\% + 47{,}61\% = 97{,}23\%$



Ejemplo 4. Determinar el área de la región bajo la curva, que está comprendida entre los puntajes de z iguales a 1,82 y 2,69.

También en este caso, es necesario expresar la región cuya área se está buscando en términos de dos regiones cuyas áreas sea posible buscar en la tabla.



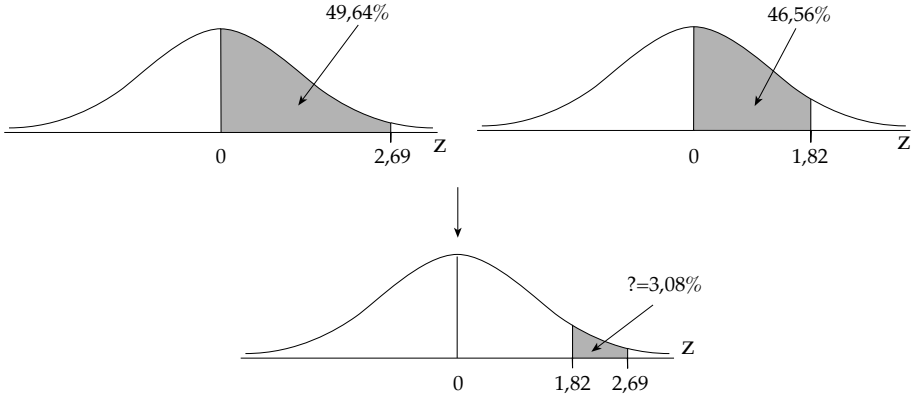
(Recuerde que dichas regiones siempre están comprendidas entre 0 y un valor positivo de z). El área de la región que aquí nos interesa se puede expresar como la diferencia de dos áreas. Primero, buscamos el área de la región comprendida entre 0 y 2,69 y a ella le sustraemos el área de la región comprendida entre 0 y 1,82. Luego, el área buscada es el resultado de la diferencia entre 0,4964 y 0,4656.

$$P(1{,}82 < z < 2{,}69) = P(0 < z < 2{,}69) - P(0 < z < 1{,}82)$$

$$P(0 < z < 2{,}69) = 0{,}4964$$

$$P(0 < z < 1{,}82) = 0{,}4656$$

Por tanto, $P(1{,}82 < z < 2{,}69) = 49{,}64\% - 46{,}56\% = 3{,}08\%$



En la tabla que hemos venido manejando hay involucrados dos tipos de datos: los puntajes z y las áreas asociadas con ellos. Ya se explicó cómo leer la tabla en caso de que se conozca el puntaje z y lo que se quiera encontrar sea el área asociada. Falta, entonces, hacer referencia al caso en que se conozca el

área asociada a un determinado puntaje z y lo que se quiera sea determinar dicho puntaje z. En realidad, el proceso que debe efectuarse es "la reversa" del anterior. Veamos unos ejemplos.

Ejemplo 5. Se sabe que el área de una región, bajo la curva, que está comprendida entre 0 y un cierto puntaje z (mayor que 0) es 0,4744. Se quiere determinar dicho puntaje.

Si se expresa el enunciado del ejercicio usando la notación correspondiente se tiene que:

$$P(0 < z < a) = 0{,}4744$$

Para encontrar el valor de **a** se procede de la manera siguiente: primero se busca en el interior de la tabla el valor 0,4744 o el número más cercano a él. Cuando se haya encontrado, —la ubicación de dicho número en la tabla se puede considerar como la intersección de una fila y de una columna— se procede a mirar los encabezamientos de la fila y la columna (en ese orden) que determinan la ubicación de dicho número. Y esos encabezamientos permiten obtener el puntaje z buscado.

Para este caso, se tiene que el valor 0,4744 está ubicado en la fila encabezada por z = 1,9 y en la columna encabezada por 0,05. Por tanto, el valor correspondiente de z es 1,95.

Ejemplo 6. Se sabe que el área de una región, bajo la curva, que está comprendida entre un cierto puntaje z (menor que 0) y 0 es 0,4744. Se quiere determinar dicho puntaje.

Si se expresa el enunciado del ejercicio usando la notación correspondiente se tiene que:[22]

$$P(-a < z < 0) = 0{,}4744$$

Ya se ha mencionado en repetidas ocasiones que la tabla sólo tiene valores positivos de z. Por tanto, en este caso el valor de **-a** se busca de manera indirecta, utilizando para ello el hecho de que la curva es simétrica con respecto de la vertical que pasa por la media. Se sabe que el área comprendida entre **-a** y 0 es la misma área que hay comprendida entre 0 y **a**. Además, se sabe que

22 El puntaje que se busca es negativo y lo estamos representando con **-a**.

dicha área es 0,4744. Utilizando, entonces, el ejemplo anterior se tiene que el valor de **a** es 1,95 y por consiguiente el valor de **-a** es -1,95.

Lo anterior puede notarse de la siguiente manera:

$$P(-a < z < 0) = 47{,}44\%$$

$$P(\mathbf{-}a < z < 0) = P(0 < z < a)$$

Por tanto, $P(0 < z < \mathbf{a}) = 47{,}44\%$ y de ahí se obtiene que $a = 1{,}95$.

a. Encuentre los siguientes números:

- $P(0 < z < 2{,}8)$
- $P(-1{,}6 < z < 0)$
- $P(-1{,}6 < z < 2{,}8)$
- $P(1{,}2 < z < 2{,}5)$
- $P(z < -2{,}3)$
- $P(z < 1{,}9)$
- $P(z > 2{,}19)$

b. Encuentre el valor de **a**, si se sabe que:

- $P(0 < z < a) = 0{,}2580$
- $P(-a < z < a) = 0{,}4582$
- $P(-x < z < x) = 0{,}3830$
- $P(-y < z < x) = 0{,}8830$
- $P(-x < z < 0) = 0{,}3830$

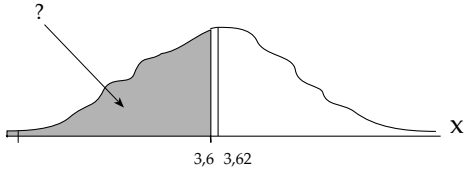
Para terminar: de vuelta a los problemas

Al iniciar este capítulo, en la sección “Motivación” se plantearon dos problemas cuya solución está aún pendiente. Ya tenemos las herramientas necesarias para resolverlos y por tanto vamos a solucionarlos a continuación.

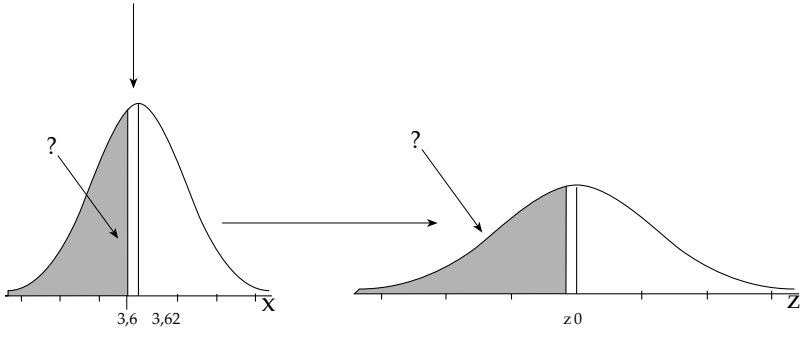
Problema 1. Se sabe que el promedio ponderado en su universidad se distribuye normalmente, con media igual a 3,62 y desviación estándar igual a 0,34. Si su promedio ponderado es 3,6, determine el porcentaje de alumnos cuyo promedio ponderado es inferior al suyo.

En realidad no se conoce la distribución de la variable en la población, pero como se sabe que tal distribución es aproximadamente normal, entonces es válido tomar como modelo para el caso, la distribución normal estándar.

Primero. Conviene hacer una gráfica que represente la situación planteada en el enunciado del problema, lo cual incluye localizar en la gráfica el valor de la media, el valor particular de la variable y además sombrear la región bajo la curva, cuya área es el dato que se debe averiguar.



Distribución real del promedio ponderado
(suposición)



Distribución "ideal" del promedio Distribución normal estándar

Segundo. Se hace la gráfica del modelo sobre el cual se va a resolver el problema —la distribución normal estándar—; en esa gráfica se localiza la media, el puntaje z asociado al valor 3,6 y se sombrea la correspondiente región bajo la curva.

Tercero. Se determina el valor del puntaje z:

En la distribución de la variable, el valor de interés en este caso es 3,6. ¿Cuál es el puntaje z asociado a dicho valor? Para responder a esa pregunta, hay que considerar que la distancia entre dicho valor y la media, (0), --medida en unidades de desviaciones estándar-- debe ser igual a la distancia que hay entre la media de la distribución de la variable, (3,62) y el valor 3,6 de la variable, --distancia, medida también en unidades de desviación estándar--. Por tanto, se tiene que:

$$3,6 - 3,62 = -0,02$$

("distancia" del valor específico de la variable a la media)

$$-0,02 \ / 0,34 = -0,05882$$

("distancia" del valor específico de la variable a la media, medida en unidades de desviación estándar)

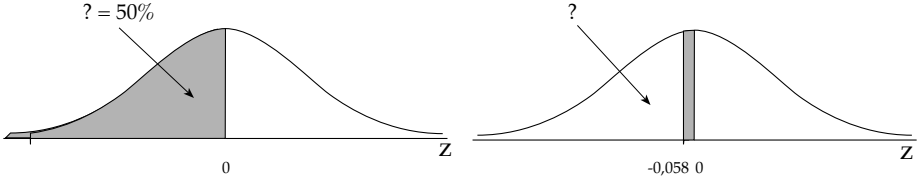
Por tanto, el puntaje z que se buscaba es z = -0,05882

Cuarto. Se emplea la tabla para determinar el área de la región sombreada en la distribución normal estándar. Dicho en otras palabras, lo que se quiere determinar es:

$$P(z < -0,05882)$$

Sin embargo, dicha área no se encuentra directamente en esa tabla. Para calcular el área deseada puede hacerse lo siguiente: la situación que debe resolverse puede replantearse es:

$$P(z < 0) - P(-0,05882 < z < 0)$$

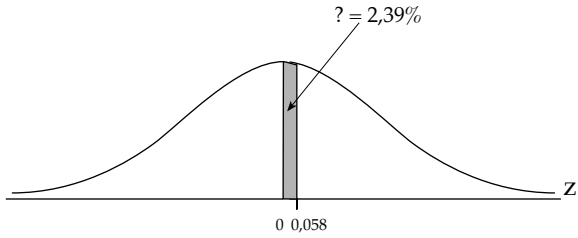


Debe, entonces, determinarse el valor de P(-0,05882 < z < 0), pero como ese valor no aparece en la tabla, debe buscarse el área de la región que está comprendida entre 0 y 0,05882 y ese valor sirve, pues las regiones son simétricas y por tanto tienen la misma área.

$$P(0 < z < 0,058) = 0,0239 \text{ (aproximadamente)}$$

Por tanto,

$$P(-0{,}058 < z < 0) = 0{,}0239 \text{ (aproximadamente)}$$

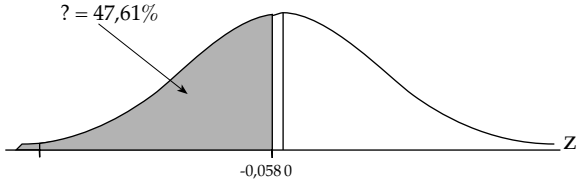


Y, para determinar $P(z < -0{,}058)$, calculamos:

$$P(z < 0) - P(-0{,}05882 < z)$$

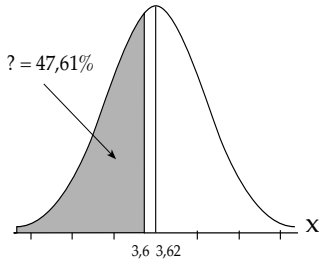
$$0{,}5 - 0{,}0239 = 0{,}4761 = 47{,}61\%$$

De lo anterior se concluye que el área de la región sombreada corresponde a un 47,61% del área total bajo la curva.



Quinto. Como el área se distribuye siempre de la misma manera en las distribuciones normales, entonces se deduce que el área de la región sombreada en la gráfica de la distribución de la variable, también es 47.61% del área total bajo la curva.

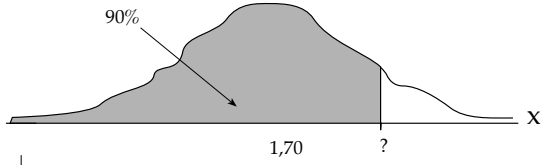
Sexto. Como el área bajo la curva se puede interpretar como la proporción de observaciones, entonces se tiene que el porcentaje de alumnos de la universidad cuyo promedio ponderado es inferior a 3,6 es 47,61%



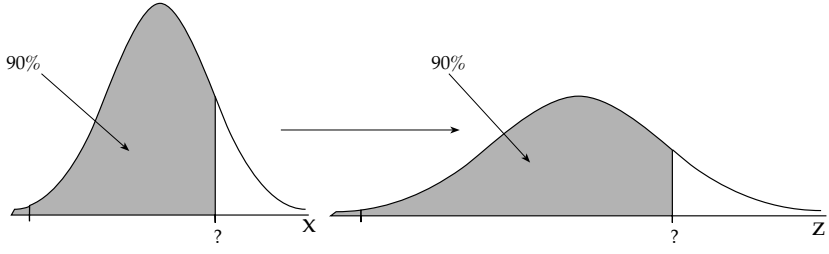
Distribución de la variable

Problema 2. Se sabe que la estatura de los estudiantes varones de su universidad se distribuye normalmente, con media igual a 1,70 metros y desviación estándar igual a 0,04 metros. ¿Cuál es la estatura mínima que debe tener un estudiante varón de la universidad, para poder pertenecer al equipo de baloncesto, si se quiere que quienes conformen el equipo tengan una estatura superior a la del 90% de la población?

En realidad no se conoce la distribución de la variable en la población, pero como se sabe que tal distribución es aproximadamente normal, entonces es válido tomar como modelo para el caso, la distribución normal estándar.



Distribución real de la estatura
(suposición)



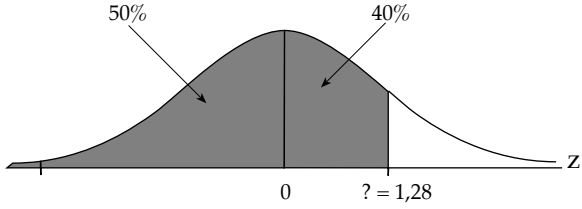
Distribución "ideal" de la estatura

Distribución normal estándar

Primero. Conviene hacer una gráfica que represente la situación planteada en el enunciado del problema, lo cual incluye localizar en la gráfica el valor de la media, sombrear la región bajo la curva, cuya área se conoce y marcar el valor de la variable que se debe averiguar.

Segundo. Se hace la gráfica del modelo sobre el cual se va a resolver el problema; en esa gráfica se localiza la media, se sombrea la región bajo la curva, que corresponde a la región sombreada en la primera gráfica, y se localiza el puntaje z asociado al valor de la variable que se debe averiguar.

Tercero. Puesto que se conoce el área de una región, se busca en la tabla el valor del puntaje z:

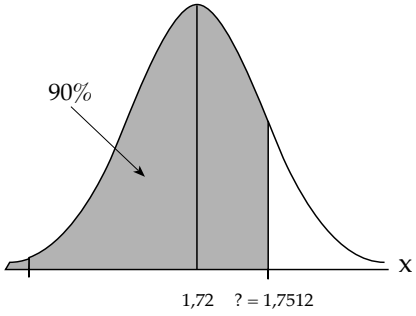


Porcentaje de área bajo la curva 90% (50% + 40%, el 50% corresponde al área de la región a la izquierda de la media 0, y el 40% restante es el área bajo la curva, entre 0 y el valor z (desconocido); esta última área es la que se busca en el interior de la tabla, para así poder determinar el valor z).

$$P(0<z<z_0)=0{,}40 \rightarrow z_0=1{,}28$$

Cuarto. En el modelo se tiene toda la información involucrada en el problema; entonces es necesario, pasar del modelo a la distribución de la variable estatura. Para ello, en la fórmula para estandarizar, se dan los valores conocidos y solucionando la ecuación que queda planteada, se encuentra el valor desconocido de la variable.

$$(?-1{,}70)/\ 0{,}04= 1{,}28 \rightarrow ?=1{,}7512$$



Distribución de la variable

Quinto. La estatura mínima que debe tener un estudiante varón para poder pertenecer al equipo de baloncesto es 1,7512 metros.

A practicar

1.- Para cada una de las distribuciones de frecuencia descritas a continuación, establezca su forma.

a. La estatura de los estudiantes de la universidad.

b. La edad de los estudiantes de este curso.

c. El primer dígito del número telefónico de todos sus amigos residentes en Bogotá.

d. El número de caras observadas al lanzar, simultáneamente, 5 monedas un número *muy grande* de veces.

e. El ingreso mensual de todos los empleados de una empresa.

f. El tiempo que usted gasta para llegar de su casa a la universidad cada mañana.

g. El número obtenido en la cara superior de un dado normal al lanzarlo un número *muy grande* de veces.

h. El resultado obtenido de sumar los puntos de las caras superiores de dos dados normales al lanzarlos simultáneamente un número *muy grande* de veces.

i. El número de veces que aparece cada vocal (ordénelas alfabéticamente) en esta oración.

2.- Enumere las características de la distribución normal.

3.- Suponga que la estatura de todos los estudiantes de su universidad tiene una distribución aproximadamente normal.

a. Determine el porcentaje de la población cuya estatura está comprendida entre el valor de la estatura promedio y el valor de la estatura que está a 1,5 unidades de desviación estándar por encima de la estatura promedio.

b. Determine el porcentaje de la población cuya estatura está comprendida entre el valor de la estatura que está a 2,25 unidades de desviación estándar por debajo de la media y el valor de la estatura que está a 2,75 unidades de desviación estándar por encima del valor de la estatura promedio.

c. Determine el porcentaje de la población cuya estatura es inferior al valor de la estatura que está a 1,75 unidades de desviación estándar de la media y por debajo de ella.

4.- Los pesos de 1.500 estudiantes varones están normalmente distribuidos con media igual a 66 kilos y desviación estándar igual a 6 kilos.

a. Determine el número de estudiantes del grupo que tienen peso igual o inferior a 51 kilos.

b. Determine el número de estudiantes del grupo que tienen peso entre 54 y 81 kilos.

c. Si se selecciona al azar, un estudiante de ese grupo, ¿qué probabilidad hay de que él tenga peso igual o superior a 72 kilos?

5.- Las calificaciones obtenidas por 200 alumnos en un examen de historia se distribuyen normalmente con media igual a 3,7 puntos y desviación estándar igual a 0,5 puntos.

a. Determine el porcentaje de alumnos que obtuvo nota inferior a 3,0 puntos.

b. Determine cuántos alumnos obtuvieron calificación entre 2,7 puntos y 4,45 puntos.

c. ¿Qué calificación debió obtener un alumno que *estuvo mejor* que el 98,78% del grupo?

6.- Mediante algunos estudios se ha establecido que el consumo de gasolina de los carros medianos se distribuye normalmente con un consumo promedio de 26 kilómetros por galón y con una desviación estándar de 4 kilómetros por galón.

a. Determine el porcentaje de carros medianos que hacen 34 o más kilómetros por galón.

b. Determine cuántos kilómetros por galón debe recorrer un nuevo carro mediano, si se quiere que éste tenga mejor rendimiento que el 95% de los autos medianos existentes.

7.- Las calificaciones de un curso están distribuidas normalmente con media igual a 3,8 y desviación estándar igual a 0,75. 10% de los estudiantes, los mejores, reciben mención de honor y 15% los peores, pierden el curso. Determine la calificación mínima para:

a. Recibir mención de honor.

b. Aprobar el curso.

8.- Las ventas diarias de un almacén tienen una distribución aproximadamente normal con media igual a $60.000 diarios y desviación estándar igual a $5.000 diarios.

a. ¿Cuál es la probabilidad de que un día cualquiera, el almacén haga ventas por más de $72.000?

b. El almacén debe tener por lo menos $49.000 en ventas diarias para poder cubrir sus costos. ¿Cuál es la probabilidad de que no pueda cubrir sus costos un día determinado?

9.- A raíz de las medidas tomadas por el gobierno para bajar el nivel de inflación en la economía colombiana, a través del incremento en el nivel de ahorro de las familias y del consiguiente aumento en la inversión y en la producción, un estudiante de Ciencia Política emprendió un estudio en marzo de 1991 con el propósito de conocer qué parte del ingreso de las familias de su barrio (estrato medio alto) se destinaba para el ahorro y con ello determinar alrededor de qué medida de tendencia central se agrupaba el nivel de ingreso de las familias de la zona. Para realizar su estudio, determinó el número de manzanas de su barrio (60 en total), las enumeró y de algunas de ellas escogió al azar una familia a la cual entrevistó. En resumen: escogió, al azar, una familia de las manzanas #1, # 3, # 5, # 7, etc., de manera que al final conformó una muestra de 30 familias residentes en su barrio, las visitó y encontró que la cantidad promedio destinada mensualmente al ahorro era de $132.800 con una desviación estándar igual a $3.225. Además, sabe que la variable se distribuye normalmente.

a. ¿Qué variable se está midiendo y de qué tipo es?

b. ¿Cuál es el objetivo de la investigación?

c. ¿Qué porcentaje de las familias encuestadas destinan al ahorro entre $131.000 y $140.000 mensualmente?

d. Suponga que de la muestra obtenida se quiere seleccionar al azar una familia que destine para el ahorro más de $142.000. ¿Cuál es la probabilidad de lograr tal suceso? Explique detalladamente el proceso empleado para solucionar el problema.

e. Después de que el estudiante tomó la muestra y la describió, decidió ampliar su investigación sobre las familias que destinan menos cantidad de sus ingresos para el ahorro. En esta ocasión, sólo visitó el 5% de las familias de la muestra que menos dinero destinan mensualmente para el ahorro. Determine cuánto es el máximo ahorro mensual que hacen dichas familias.

f. Comente si la siguiente afirmación es verdadera o es falsa y justifique su respuesta.

> El 68% de los residentes en el barrio destina mensualmente al ahorro entre $136.000 y $199.000.

10.- Se realizó un estudio en la Universidad X para estimar la edad promedio de los estudiantes de dicha universidad que sufragaron en las elecciones pasadas para elegir alcalde. Tomaron una muestra de 300 estudiantes de la población y encontraron que la variable se distribuye normalmente, la edad promedio es 24 años y tiene desviación estándar de 2,2 años. Uno de los investigadores hace las siguientes afirmaciones con respecto a la muestra:

- Menos del 1% de los estudiantes considerados en la muestra tienen entre 18 y 19 años.
- El 20% de los estudiantes "más viejos" considerados en la muestra tienen más de 25 años.
- Aproximadamente 279 de los estudiantes considerados en la muestra tienen entre 20 y 28 años.
- Si se selecciona, al azar, un estudiante de esa muestra, la probabilidad de que éste tenga menos de 19 años o más de 25 es 0,338.

a. Determine cuál es la población y cuál es la muestra de estudio para el grupo de politólogos.

b. ¿Cuál es el objetivo de la investigación?

c. ¿Cuál es la variable de interés y de qué tipo es?

d. Con respecto a las afirmaciones hechas acerca de la muestra diga si son verdaderas o no y en cada caso justifique detalladamente su respuesta.

11.- Una psicóloga aplicó una serie de pruebas proyectivas a los niños de tercero elemental de un colegio, con el fin de conocer determinados aspectos psicológicos de ellos. Para realizar dichas pruebas, los niños debían pintar el cuerpo humano, su familia y contar algo acerca de lo que habían pintado. La psicóloga interpretó la realización de cada niño y lo que dijo, y asignó un puntaje entre 0 y 5 de acuerdo a ciertos items ya establecidos. Una vez terminadas las pruebas, promedió los puntajes y encontró que la distribución de éstos es aproximadamente normal con media $\bar{x} = 3{,}6$ >y desviación estándar $s = 0{,}5$.

a. Si se considera que los niños con un puntaje menor o igual a 2,5 tienen problemas psicológicos, calcule el porcentaje de niños bajo prueba que presentan dichos problemas.

b. Si se considera que los niños con un puntaje mayor o igual a 4,6 presen-

tan indicios de que son superdotados, ¿qué porcentaje de los niños bajo prueba presentan estos indicios?

c. Determine entre qué par de puntajes debe estar ubicado el promedio obtenido por un alumno que haga parte del "montón" conformado por el 80% del total de la muestra (es decir, se excluye el 10% de los que tienen el peor puntaje y también se excluye el 10% de los que tienen el mejor puntaje).

12.- El enfoque funcionalista del lenguaje se centra en el análisis de las funciones que cumple el lenguaje dentro del proceso de comunicación. Los políticos, personas con gran capacidad oratoria y comunicativa, tienen que manejar las diferentes funciones, especialmente la emotiva y poética, para lograr que sus palabras impacten y convenzan a su auditorio. Preocupado por el importante papel del discurso en la contienda electoral, un asesor de campaña de un candidato al Senado de la República adelantó una investigación que pretendía estudiar el poder de convencimiento del candidato. Para esto, encuestó a 20 personas después de las cinco primeras intervenciones del candidato, en plaza pública, pidiendo a los encuestados que calificaran sobre 5,0 qué tan verdaderas les habían parecido las afirmaciones del candidato durante el discurso. Se encontró como media de la muestra 3,7 y como desviación estándar 0,35.

a. ¿Cuál es la muestra de estudio?

b. Atendiendo a los valores de la media y de la desviación estándar encontrados en la muestra, ¿cree usted que el problema pueda trabajarse usando el modelo de la distribución normal? Justifique su respuesta.

Para contestar las siguientes preguntas suponga que la variable se distribuye normalmente.

c. ¿Cuál es la probabilidad de que una persona no le crea al candidato? (Piense en la calificación mínima que usted debe obtener para aprobar un quiz calificado sobre 5,0.)

d. ¿Qué porcentaje de personas sí le creen al candidato?

e. ¿Cuál es la calificación que indica que el 80% de las personas creen en el candidato?

f. ¿Entre qué par de valores se encuentra el 70% de las calificaciones si se sabe que del 30% restante, 17% no le creen y 13% sí?

13.- Uno de los fenómenos de estudio más comunes dentro de los enfoques normativos de la sociolingüística es el uso de dos lenguas en un mismo contexto, es decir, el bilingüismo. Muchas investigaciones han mostrado "la relación entre la conducta verbal y una variedad de factores psicológicos y sociales",[23] e incluso sociolingüistas como Joshua Fishman han desarrollado la teoría de que, en comunidades bilingües con diglosia,[24] los miembros de tal comunidad tienden a utilizar un discurso que expresa intimidad, solidaridad, espontaneidad e informalidad en "dominios lingüísticos"[25] informales como la familia o el grupo de amistad; mientras que en dominios más formales como la educación, la ocupación o la religión, las personas tienden a utilizar un discurso que involucra diferencias de status, ritos o formalidad.[26]

Con el fin de estudiar las particularidades de la comunidad bilingüe de puertorriqueños en Nueva York, se adelantó una investigación que, por medio de pruebas aplicadas a 215 miembros de la comunidad, medía la cantidad de inglés que la persona usaría en determinado dominio. Las personas debían calificar de 0 a 5 (donde 0 indica que no hay uso del inglés y 5 indica que la comunicación es toda en inglés) el lenguaje que usarían en una situación hipotética presentada. Los resultados de las pruebas fueron:

	Dominio				
	familia	**amistad**	**religión**	**educación**	**trabajo**
media, $\bar{x}$	2,26	2,70	4,09	4,83	4,50
desviación estándar, s	1,15	1,22	1,19	0,79	0,67
tamaño de la muestra, n	215	215	215	215	215

23 Greenfield, Lawrence. "Situational Measures of Normative Language Views in Relation to Person, Place end Topic among Puerto Rican Bilinguals". *Advances in the Sociology of Language, Joshua Fishman ed. Paris, Mouton Publishers, 1972, p. 17.*

24 La diglosia es la diferenciación que las personas bilingües hacen sobre cuándo, dónde y con quién se debe hablar uno u otro idioma.

25 Un *dominio lingüístico* es el contexto institucional dentro del cual el uso habitual del lenguaje tiene lugar. En esta investigación se identificaron cinco dominios: familia, amistad, religión, educación y trabajo, ordenados según su grado de formalidad.

26 *Ibid.*, pp. 16-20.

a. ¿Cuál es el problema de estudio?

b. En sus palabras, exponga el marco teórico de la investigación.

c. Haciendo uso de la información presentada, determine si las siguientes afirmaciones son verdaderas:

- La probabilidad de que una persona hable todo en inglés aumenta a medida que aumenta la formalidad del dominio.
- El porcentaje de observaciones por encima de la calificación que indica igual cantidad hablada de inglés y español aumenta con la formalidad del dominio.
- El porcentaje de observaciones contenidas a dos desviaciones estándar de la media ($\bar{x} + 2s$) es igual para todos los dominios.

d. Tomando en cuenta las respuestas anteriores, ¿estas distribuciones son aproximadamente normales? Justifique su respuesta.

e. Proponga una manera de arreglar la información presentada de tal forma que el problema sí pueda ser resuelto utilizando la distribución normal.

14.- En el examen de estadística hecho a un grupo de 150 alumnos, la calificación promedio fue 78 y la desviación estándar 8. En el examen final de inglés realizado al mismo grupo de estudiantes, la calificación promedio fue 73 y la desviación estándar fue 7,6. Para responder las siguientes preguntas, suponga que la distribución de las calificaciones en ambas materias es normal.

a. Si Andrés es estudiante de tal curso y obtuvo 75 en estadística y 71 en inglés, ¿en cuál de los dos exámenes tuvo una calificación relativa superior? Explique.

b. Si en el examen de inglés, una persona *pasa* con una calificación de 60 o más, ¿cuántos alumnos del curso *pasaron*?

c. Si en el examen de estadística, la calificación de un alumno está por encima de la calificación del 90% del grupo, ¿cuál es dicha calificación?

Estadística inferencial

Introducción

En este capítulo se tratarán generalidades referentes a la estadística inferencial. En primer lugar, se presentarán cuatro problemas cuyas soluciones atañen a esa rama de la estadística. La intención que se tiene al proponer dichos enunciados es identificar algunos tipos de preguntas que se pueden responder con la ayuda de la estadística (no se solucionarán en este capítulo). En la siguiente sección se responderá a preguntas tales como: qué es inferir, por qué es necesario hacer inferencia, cuáles son los procesos de inferencia utilizados, qué tan válidos son los procesos de inferencia, y finalmente se definirá el concepto de distribución muestral. En la tercera sección se hará referencia a la distribución muestral de medias. Y en la última sección se trabajará la distribución muestral de diferencias de medias.

Motivación

A continuación se enuncian cuatro problemas. Léalos, tratando de identificar qué se quiere hacer en cada uno de ellos. Además, imagínese que es usted quien está enfrentado a cada una de esas situaciones problemáticas y por tanto es usted quien debe decidir qué información se requiere para dar solución al problema.

Problema 1. Con el fin de revisar algunas cláusulas de las pólizas de seguros de vida, un corredor de seguros quiere determinar la edad promedio de muerte de adultos que fallecen de manera natural en la ciudad X.

Problema 2. En un centro de estética, durante los últimos seis meses, se han estado empleando dos tratamientos diferentes para reducir de peso (T_1 y T_2). El tratamiento T_1 se ha aplicado a un grupo G_1, mientras que el tratamiento T_2 se

ha aplicado a un grupo G_2. Ambos grupos están formados por adultos cuyas edades oscilan entre 25 y 35 años, que tienen problemas de obesidad. El tratamiento T_2 es sustancialmente más costoso que el T_1. El médico del centro quiere determinar entre qué par de valores se puede esperar que esté la diferencia en los pesos medios rebajados después de los tratamientos para tomar decisiones hacia el futuro con respecto al tratamiento que debe ofrecer el centro.

Problema 3. El productor de cigarrillos de la marca A afirma que el contenido medio de nicotina por cigarrillo es de 0,30 miligramos. Un grupo de médicos quiere verificar si es posible aceptar como cierta la afirmación hecha.

Problema 4. Un profesor de pre-escolar conoce dos métodos para enseñar a leer y sospecha que el método A produce mejores resultados que el método B. El quiere verificar su hipótesis.

a. ¿Qué se pide hacer en cada uno de los problemas? Sea tan explícito como le sea posible.

b. Compare los enunciados de los problemas 1 y 3, en términos de lo que se pide realizar en cada uno de ellos.

c. Compare los enunciados de los problemas 1 y 2, en términos de lo que se pide realizar en cada uno de ellos.

d. Si usted quisiera abordar las situaciones problemáticas planteadas, ¿que información debería tener? Para cada caso, sea tan explícito como le sea posible.

e. ¿Se da usted cuenta de la presencia del azar en las situaciones problemáticas expuestas? Explique su respuesta.

f. Proponga un problema (ojalá que tenga que ver con su carrera) cuyo enunciado sea del mismo tipo que el de alguno de los dos primeros problemas planteados.

g. Proponga un problema (ojalá que tenga que ver con su carrera) cuyo enunciado sea del mismo tipo que el de alguno de los dos últimos problemas planteados.

Al analizar los enunciados de los problemas se encuentran algunas semejanzas y también algunas diferencias; precisamente a través de las diferencias y las semejanzas que se detecten, intentaremos lograr una descripción general de los problemas que centrarán nuestro interés en lo que resta del texto. Miremos con algún detalle los diferentes enunciados.

En el primer problema se quiere hacer una generalización sobre la población de adultos que fallecen, de manera natural, en la ciudad X; el aspecto de interés que se está cuantificando es la edad que tiene la persona a la hora de su muerte; lo que se quiere hacer es determinar la edad promedio de muerte en la correspondiente población. Para responder a esta pregunta es necesario contar con una muestra aleatoria extraída de la población de datos.

En el segundo problema se quiere hacer una generalización sobre la población de adultos de edades entre 25 y 35 años, que tienen problemas de obesidad y que llegan al centro de estética mencionado con la intención de seguir tratamiento para reducir de peso; el aspecto de interés que se está cuantificando es el peso rebajado por quienes siguen tratamiento para reducir de peso; lo que se quiere hacer es determinar la diferencia en los pesos medios rebajados, de quienes conforman la población. Para responder a esta pregunta es necesario contar con dos muestras aleatorias e independientes entre sí: una, (M_1), que registre el peso rebajado por las personas del grupo G_1 y, otra muestra, (M_2), que registre el peso rebajado por las personas del grupo G_2.

En el tercer problema se quiere hacer una generalización sobre la población de cigarrillos de la marca A; el aspecto de interés que se está cuantificando es la cantidad de nicotina presente en cada cigarrillo; en este problema ya no interesa determinar un cierto valor de la población; ahora se tiene una hipótesis acerca del contenido medio de nicotina de un cigarrillo de la marca A y se desea aceptarla o rechazarla. Para responder a esta pregunta es necesario contar con una muestra aleatoria obtenida de la población de datos correspondiente.

El problema cuarto está vagamente definido pues, por ejemplo, no se especifica la población; sin embargo, se quiere hacer una generalización sobre la población (cualquiera que ella sea); el aspecto en el que se tiene interés es la calidad de los métodos A y B; y al igual que en el problema anterior no se quiere determinar un valor de la población de datos, sino más bien decidir si se puede aceptar o no una hipótesis con respecto a la diferencia de calidad de los dos métodos. Para responder a esta pregunta es necesario contar con dos muestras aleatorias obtenidas de la población que se defina y a la cual se podría aplicar la generalización que se haga.

De las cuatro situaciones problemáticas también podemos afirmar que, aunque no se haga explícito en el enunciado, se dan en un contexto específico que incluye una experiencia previa, la observación de ciertos hechos, unas intuiciones con respecto al problema de interés, y además la necesidad de tomar decisiones.

Por otra parte, en las situaciones descritas se hace evidente la imposibilidad de trabajar exhaustivamente con toda la población de interés; por tanto se hace necesario seleccionar una muestra que represente a la población. En esta selección está presente el azar.

Las anotaciones hechas con respecto a los enunciados se pueden resumir así:

- La idea central, en los cuatro casos, es hacer generalizaciones sobre el comportamiento de una población en un determinado aspecto.
- El aspecto alrededor del cual se quiere hacer la generalización es un aspecto observable y medible cuantitativamente; dicho en otras palabras, se trata de una variable cuantitativa.

En términos generales podemos aceptar que los puntos anteriores caracterizan de manera aceptable los problemas en que centraremos nuestra atención en los próximos capítulos. Sin embargo, podemos ser más específicos con respecto al tipo de generalización que se pretende realizar sobre la población. Básicamente, se van a desarrollar dos tipos de generalización:

- Estimación del valor de un parámetro.[27]
- Prueba de hipótesis sobre el valor de un parámetro.

En los próximos capítulos se explicará en qué consiste cada uno de estos dos procedimientos; por ahora, basta con saber que lo que se pide en los dos primeros problemas es la estimación de un parámetro, y lo que se pide en los otros dos problemas es la realización de una prueba de hipótesis sobre un parámetro.

27 Recuérdese que cualquier medida que describa el comportamiento de una variable en una población se denomina parámetro. Por ejemplo, la media aritmética de una población de datos, la desviación estándar de la misma, etc. son parámetros de tal población. En cambio, cualquier medida que describa el comportamiento de una muestra se denomina estimador de parámetro o estadístico. Por ejemplo, la media aritmética de una muestra, la desviación estándar de una muestra, etc. son estadísticos. Tanto los parámetros como los estimadores se refieren a conceptos, en tanto que se habla de *estimativos* cuando se hace referencia al valor numérico de un estimador.

Algunos conceptos fundamentales

En esta sección vamos a explicar algunos conceptos que son fundamentales para entender **en qué consiste** la tarea de la estadística inferencial y **por qué** los procedimientos que utiliza la estadística inferencial son válidos.

Como se dijo anteriormente, los dos primeros problemas —el del corredor de seguros y el del centro de estética— exigen la estimación de parámetros, y los otros dos —el del grupo de consumidores y el del profesor de pre-escolar— exigen la validación de una hipótesis. Tenga en cuenta eso para responder las siguientes preguntas.

a. ¿Qué cree usted que significa en estadística la expresión "estimar el valor de un parámetro"? ¿Qué diferencia encuentra entre "calcular" y "estimar"? Explique su respuesta.

b. ¿Qué cree usted que significa en estadística la expresión "validar una hipótesis"? ¿Qué diferencia encuentra entre "demostrar" y "validar"? Explique su respuesta.

c. ¿Qué cree usted que significa en estadística la expresión "inferir"?

Si nos remitimos al diccionario de la lengua española se encuentra que *inferir* y *deducir* son palabras sinónimas y su significado es "sacar consecuencias de un principio, proposición o supuesto". Sin embargo, la connotación de tales palabras en estadística es un tanto diferente. Para hacer claridad en este aspecto vamos a citar algunos apartes del artículo titulado "Estadística" de Warren Weaver que fue publicado en *Matemáticas en el mundo moderno*.

> *Existen dos formas principales de pensamiento lógico, la deducción y la inducción. La primera se debe principalmente a los griegos, que fueron los primeros en ver claramente la gran potencia de proponer axiomas o hipótesis generales y deducir de ellos una ordenación útil de proposiciones implicadas por ellos. El pensamiento inductivo, [...] no comenzó a constituir una herramienta sistemática del hombre hasta la última parte del siglo XVIII. La inducción procede en la dirección opuesta a la deducción. Partiendo de hechos experimentales, nos conduce a inferir conclusiones generales.*

> *El razonamiento deductivo es tajante y absoluto. Sus injerencias específicas se siguen inevitablemente de hipótesis generales. El razonamiento inductivo, al contrario, es una inferencia incierta. Los hechos concretos y especiales de la experiencia, a partir de los cuales comienza el razonamiento inductivo, generalmente no conducen inexorablemente a conclusiones generales categóricas. Más bien conducen a juicios que se refieren a la plausibilidad de diversas conclusiones generales.*
>
> *[...] la estadística es el nombre de la ciencia y del arte que trata de la inferencia incierta, la cual usa los números para obtener algún conocimiento acerca de la naturaleza y de la experiencia.*
>
> *[...] Lo importante del razonamiento inductivo se basa en el hecho de que, dejando a un lado excepciones triviales, los sucesos y los fenómenos de la naturaleza son demasiado multiformes, demasiado numerosos, demasiado extensos o demasiado inaccesibles para permitir una observación completa. [...] No podemos medir los rayos cósmicos en todas partes y en cada instante. No podemos ensayar un nuevo medicamento en todas las personas. [...] Así, hemos de contentarnos con muestras. Las medidas obtenidas en cada experimento científico constituyen una muestra del conjunto ilimitado de mediciones que resultarían si uno realizase el mismo experimento una y otra vez indefinidamente. Casi siempre se interesa uno en la muestra solamente en cuanto que es capaz de revelar algo acerca de la población de la cual procede.*

De manera que lo que en este texto vamos a entender como inferencia es un procedimiento lógico —basado en la inducción y no en la deducción— que permite llegar a conclusiones generales pero no categóricas; dicho más exactamente, se obtienen "juicios que se refieren a la plausibilidad de diversas conclusiones generales".

Ahora bien, la inferencia estadística se lleva a cabo bajo dos formas: la estimación y la validación de hipótesis. Veamos a grandes rasgos en qué consiste y qué alcance tiene cada uno de esos procedimientos.

La estimación de un parámetro permite determinar, **con alguna probabilidad de acertar**, un intervalo en el cual es posible encontrar tal parámetro. Se ha dicho **es posible** y no **es seguro**. La afirmación anterior enfatiza lo dicho con respecto a que la estadística se encarga de la inferencia incierta. Por ejemplo, si se quiere estimar el tiempo medio que emplea una persona que vive en determinada zona de la ciudad para ir de su casa al lugar de trabajo, la forma de proceder consiste en tomar una muestra aleatoria y representativa de la población para obtener la información a partir de la cual se va a inferir; después de realizar la estimación (esto se estudiará en el próximo capítulo) se

tendrá una afirmación del siguiente estilo: "se estima que el tiempo medio empleado, por una persona de la población en cuestión, para llegar del sitio donde reside al sitio donde trabaja está entre los valores a y b, con una certeza del $x\%$".

La validación de una hipótesis acerca del valor de un parámetro permite rechazar o no, **con algún margen de error** la hipótesis estadística que se ha formulado con relación a dicho parámetro. En la realidad no es posible determinar si la hipótesis formulada es cierta o no lo es; y, por tanto, la conclusión a la que se llega a través de la realización de la prueba de hipótesis es incierta; será valiosa y cercana a la realidad en la medida en que la muestra a partir de la cual se está haciendo la inferencia sea aleatoria y represente bien a la población a la que pertenece. Por ejemplo, para validar la hipótesis de que el tiempo medio que emplea una persona —que vive en una cierta zona de la ciudad— para ir de su casa al sitio donde trabaja es menor que 1,5 horas, la forma de proceder consiste en tomar una muestra aleatoria y representativa de la población de datos para obtener la información a partir de la cual se va a inferir; después de realizar la prueba de hipótesis (esto se estudiará más adelante) se tendrá una afirmación, por ejemplo, del siguiente estilo: "se puede rechazar la hipótesis con una probabilidad del $x\%$" de cometer error.

Retomemos el tema de la extracción de muestras de una población (cuya finalidad es allegar información para hacer inferencias sobre dicha población). A este respecto no parece necesario hacer una justificación elaborada para convencer al lector de la necesidad de este proceso; como bien lo señala Weaver en el artículo anteriormente mencionado, "El muestreo no es meramente conveniente. Es a menudo la única forma posible de tratar un problema". Sin embargo, es importante hacer algunos comentarios con relación a la validez del muestreo para hacer inferencia y con relación a la confiabilidad de las inferencias hechas así. Probablemente, habrá quienes crean que no es válido inferir sobre la población con base en **una** sola muestra, o piensen que la muestra debe ser "muy grande", además de representativa y aleatoria. En resumen, debemos responder de manera satisfactoria las dos preguntas:

- ¿Por qué es **válido** hacer inferencia sobre un parámetro con base en la información que arroja **una sola muestra** tomada de la población?
- ¿Qué tan **segura** es la inferencia que se puede hacer con base en la información que arroja **una sola muestra** tomada de la población?

Imagínese que puede extraer, de la población acerca de la cual quiere inferir, todas las muestras posibles de un determinado tamaño y que además hace la selección de las muestras con sustitución y con orden.[28] Suponga también que para cada una de las muestras calcula el valor que interesa para el caso, es decir, el estimativo del correspondiente estimador (por ejemplo, si se quiere inferir acerca de la media de la población, entonces se debería calcular la media de cada una de las muestras). El conjunto de valores resultantes es una distribución de datos con un valor promedio, una desviación estándar, su gráfica tiene una forma, etc. Pues bien, es en esa distribución teórica (en la realidad no es posible construirla, solamente la imaginamos) en la que se apoya la lógica de la inferencia estadística ya que el estimativo que aporta la muestra —con base en la cual se hace la inferencia— es un valor de la distribución y ella se puede modelar con alguno de los modelos estadísticos que se conocen. Es decir, aplicamos las características de un determinado modelo para concluir acerca de una distribución muestral, la cual se puede relacionar de manera precisa con la distribución de la población, siendo así posible la inferencia acerca de la población.

En las propias palabras de Weaver:

> *Es evidente que [...] nunca puede afirmar con certeza cómo es la población original, mediante un mero muestreo, porque las demás muestras variarán. Sin embargo, para una cierta clase de población y con métodos adecuados de muestreo es posible elaborar teóricamente el esquema de variación de las muestras. Este conocimiento del esquema de variabilidad de muestras da un firme apoyo. Permite considerar las muestras y obtener conclusiones acerca de la población original.*

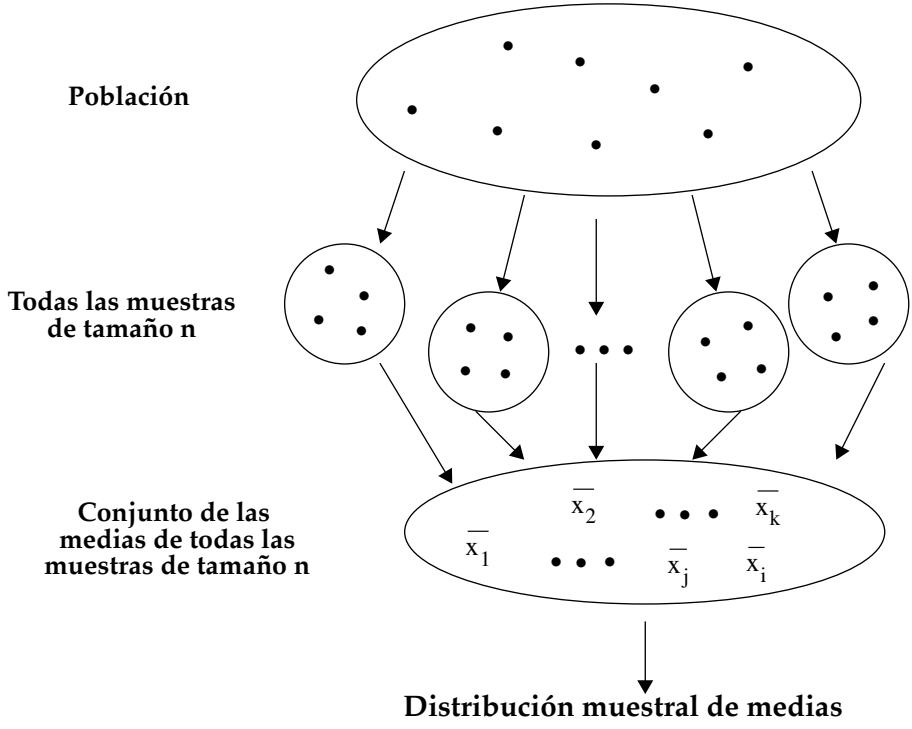
En resumen, el estimativo de la muestra con base en el cual se pretende inferir es un valor de una distribución teórica, cuyos datos tienen variabilidad y esa variabilidad sigue alguna tendencia; por decirlo de una manera muy elemental, los posibles valores que puede tomar el estadístico no varían caóticamente. Así pues, el valor del estadístico de una muestra sí permite obtener conclusiones generales que informan sobre la población. Por otra parte, el tamaño de la muestra y la técnica de muestreo empleada para obtenerla son factores decisivos en la calidad de las inferencias.

---

28 Estos dos supuestos se hacen con el fin de poder referirnos a poblaciones que aunque no sean infinitas, permitan repetir indefinidamente el muestreo; si no se hiciera el supuesto de sustitución, en el proceso de sacar muestras podría eventualmente agotarse la población.

La confiabilidad de la inferencia tiene que ver con la posibilidad de cometer error. Ninguna inferencia estadística está libre de error puesto que no se trabaja con toda la población sino con sólo una porción de ella, y de muestra a muestra hay variabilidad. Entonces, es posible llegar a tener una actitud escéptica con respecto al alcance de estos procedimientos, al pensar que si la muestra con base en la cual se está trabajando es "rara", entonces la inferencia será muy poco confiable. Aunque lo anterior puede llegar a ser cierto, es muy poco probable; precisamente, si la muestra es atípica, tendrá muy poca probabilidad de ser escogida, y en cambio, las muestras más típicas tienen más probabilidad de ser extraídas.

Una distribución teórica como la que fue imaginada anteriormente se llama *distribución muestral*.



El diagrama anterior resume el proceso que habría que desarrollar para construir una distribución muestral. Particularmente, vamos a referirnos a la construcción de la distribución muestral de medias.

> Una *distribución muestral de un estimador* está constituida por todos los valores de un estadístico dado, calculado para todas las muestras, de un mismo tamaño, que es posible extraer de una población.

Dada la importancia de las distribuciones muestrales para hacer inferencia estadística, es naturalmente obvia la necesidad de estudiarlas con algún detalle. En este texto, sólo nos interesa hacer inferencia sobre la media de una población y sobre la diferencia de medias de dos poblaciones. Por tanto, estudiaremos en detalle la distribución muestral de medias y la distribución muestral de diferencias de medias.

a. Para un caso particular en el que usted tenga que hacer inferencia sobre la media poblacional, ¿cree que el proceso de inferencia requiera la construcción de la distribución muestral de medias? Explique su respuesta.

b. Explique en sus palabras el significado de la siguiente afirmación:

> El proceso para inferir sobre la media de una población se apoya en la distribución muestral de medias.

Distribución muestral de medias

Aunque se ha dado ya una definición del concepto de distribución muestral, del cual se debería deducir el concepto de distribución muestral de medias, parece necesario aclarar este último concepto mediante un ejemplo, que servirá además para poner en evidencia el llamado *teorema del límite central*.

Vamos a proponer entonces una población hipotética para construir y estudiar, tres distribuciones muestrales del estimador $\bar{x}$.

En el ejemplo que vamos a trabajar conocemos la población. En la realidad, un hecho como éste no ocurre, pues en tal caso no habría necesidad de inferir. El

ejemplo se escogió así porque facilita relacionar las características de la población con las de la distribución muestral, y de esa manera será posible verificar resultados de la teoría que se van a emplear frecuentemente y cuya demostración está fuera del alcance de este texto.

Suponga que la población es P = {1, 2, 3, 5} y que representa el tiempo (en horas diarias) que cada uno de un grupo de cuatro estudiantes de la universidad dedica a preparar sus tareas académicas.

a. Calcule el promedio y la desviación estándar de la población. Además, haga un dibujo que represente la distribución de la población.

Ahora vamos a observar cómo se comportan las medias de las muestras de la población. En primer lugar, vamos a trabajar con las muestras de tamaño 2.

b. Si se hace la selección de las muestras con orden y con sustitución y ellas son de tamaño 2, ¿cuántas hay?

El conjunto de muestras de tamaño 2 de la población P tiene 16 elementos, todos diferentes. Si para cada muestra se calcula su correspondiente media entonces se tiene una distribución de 16 elementos, no todos diferentes. El mínimo de esa distribución es (1), que es el promedio de {1, 1}.

c. ¿Cuál es el máximo de la distribución de medias.

d. ¿Cuáles son todos los valores posibles de las medias en este caso?

En el interior de la siguiente tabla de dos entradas se escriben las medias de las correspondientes muestras de tamaño 2.

Medias de las muestras de tamaño 2				
	1	2	3	5
1	1	1,5	2	3
2	1,5	2	2,5	3,5
3	2	2,5	3	4
5	3	3,5	4	5

media de la muestra {1, 5}

La información que da la tabla anterior se puede organizar en una tabla de distribución de frecuencias así:

Distribución medias muestrales (n=2)	
Valores de la media	**Frecuencia**
1	1
1,5	2
2	3
2,5	2
3	3
3,5	2
4	2
5	1

Hemos construido la distribución muestral de medias de tamaño 2. Esa distribución, igual que toda distribución, tiene una gráfica de una determinada forma, una media, una desviación estándar. No haga ningún cálculo: sólo piense. ¿Cree usted que la media de la distribución de la que estamos hablando coincide con la media de la población? Y, ¿cree que la desviación estándar de la población coincide con la de la distribución que hemos construido? Trate de explicar su intuición.

e. Compruebe que la media de la distribución muestral de medias (muestras de tamaño 2) es igual a 2,75 y que la desviación estándar es igual a 1,045825. Además, haga la gráfica de la distribución.

Ahora, tomemos de la población P todas las muestras de tamaño 3, con sustitución y con orden.

f. ¿Cuántas de tales muestras hay?

El conjunto de muestras de tamaño 3 de la población P tiene 64 elementos, todos diferentes. Si para cada muestra se calcula su correspondiente media entonces se tiene una distribución de 64 elementos, no todos diferentes. El mínimo de esa distribución es (1) que es el promedio de {1, 1, 1}.

g. ¿Cuál es el máximo de la distribución de medias?

h. Dé ejemplo de cinco muestras de tamaño 3, extraídas de la población P; además, para cada una de ellas obtenga la correspondiente media.

i. ¿Cuáles son todos los valores posibles de las medias en este caso?

La siguiente tabla presenta las 64 muestras posibles, junto con la correspondiente media:

#	Muestra			Promedio	#	Muestra			Promedio
1	1	1	1	1	33	3	1	1	5/3
2	1	1	2	4/3	34	3	1	2	2
3	1	1	3	5/3	35	3	1	3	7/3
4	1	1	5	7/3	36	3	1	5	3
5	1	2	1	4/3	37	3	2	1	2
6	1	2	2	5/3	38	3	2	2	7/3
7	1	2	3	2	39	3	2	3	8/3
8	1	2	5	8/3	40	3	2	5	10/3
...			...		...			...	
25	2	3	1	2	57	5	3	1	3
26	2	3	2	7/3	58	5	3	2	10/3
27	2	3	3	8/3	59	5	3	3	11/3
28	2	3	5	10/3	60	5	3	5	13/3
29	2	5	1	8/3	61	5	5	1	11/3
30	2	5	2	3	62	5	5	2	4
31	2	5	3	10/3	63	5	5	3	13/3
32	2	5	5	4	64	5	5	5	5

Al emplear la información anterior para hacer la distribución de frecuencias de las medias muestrales (de tamaño 3), se tiene:

Distribución de medias muestrales (n = 3)	
Promedio	**Frecuencia**
1	1
4/3	3
5/3	6
2	7
7/3	9
8/3	9
3	10
10/3	6
11/3	6
4	3
13/3	3
5	1

Así hemos construido la distribución muestral de medias de tamaño 3. El rango de esta distribución es igual al rango de la primera distribución muestral que hicimos, e igual al rango de la población, pero la distribución de las medias de las muestras de tamaño 3 es más densa y tiene menos dispersión.

j. Compruebe que la distribución de medias de muestras de tamaño 3, tiene promedio igual a 2,75 y la desviación estándar es aproximadamente igual a 0,853912.

k. Haga la gráfica de la distribución.

Se quiere construir ahora la distribución muestral de medias (de tamaño 4). En este caso hay 4 * 4 * 4 * 4 = 256 muestras diferentes. Si para cada una de esas muestras se obtiene la correspondiente media, entonces se genera un conjunto de 256 elementos no todos diferentes. El mínimo valor de ese conjunto es 1 y el máximo es 5.

l. Dé ejemplo de cinco de tales muestras, con la correspondiente media. Además, haga la lista de los valores que pueden ser media de alguna muestra de tamaño 4.

La siguiente es la distribución de frecuencias de las medias muestrales (de tamaño 4):

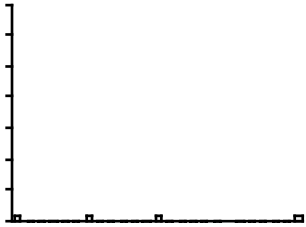
Distribución de medias muestrales (n = 4)				
Promedio	**Frecuencia**		**Promedio**	**Frecuencia**
1	1		3	31
5/4	4		13/4	24
6/4	10		14/4	22
7/4	16		15/4	12
2	23		4	10
9/4	28		17/4	4
10/4	34		18/4	4
11/4	32		5	1

La media de esta distribución es 2,75 y la desviación estándar es igual a 0,739509.

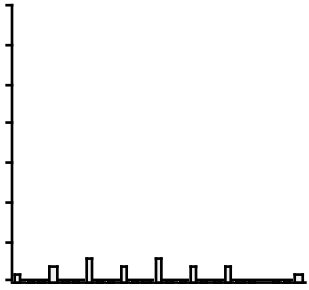
En resumen, se han construido las tres distribuciones muestrales de medias, asociadas con la población P. Las características de la población P y de las tres distribuciones muestrales se exponen a continuación:

	tamaño	**media**	**desviación estándar**
Población	4	2,75	1,479019
Distribución muestral de medias (n = 2)	16	2,75	1,045825
Distribución muestral de medias (n = 3)	64	2,75	0,853912
Distribución muestral de medias (n = 4)	256	2,75	0,739509

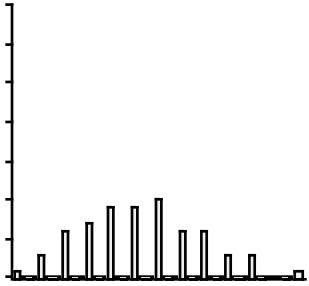
Distribución de la población



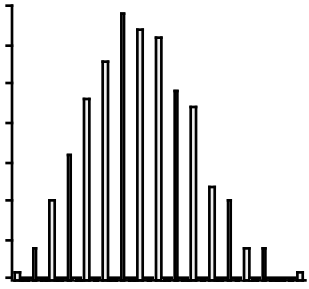
Distribución de las medias de las muestras de tamaño 2



Distribución de las medias de las muestras de tamaño 3



Distribución de las medias de las muestras de tamaño 4



m. La gráfica de la población y de las tres distribuciones muestrales se presentaron en la página anterior. Obsérvelas y tenga en cuenta la información de la última tabla para comparar el comportamiento de la población con el de cada una de las distribuciones muestrales. Además, compare el comportamiento de las tres distribuciones muestrales.

Al comparar los cuatro diagramas se observa una evolución: el de la población es uniforme y los diagramas de las distribuciones muestrales van aproximándose a la curva normal a medida que el tamaño de las muestras se aumenta. También vemos que las medias de las cuatro distribuciones coinciden y en cambio, las desviaciones estándar disminuyen a medida que aumenta el tamaño de las muestras.

Veamos cómo se relacionan la desviación estándar de la población con la desviación estándar de la distribución muestral y con el tamaño de las muestras. Observe las siguientes expresiones:

$$1,045825033 \times \sqrt{2} = 1,479019945$$

$$0,853912565 \times \sqrt{3} = 1,479019948$$

$$0,739509972 \times \sqrt{4} = 1,479019944$$

Los tres productos dan, prácticamente, el mismo resultado que es el valor de la desviación estándar de la población. En realidad, el producto entre la desviación estándar de la distribución muestral de medias y la raíz cuadrada del tamaño de las muestras es igual a la desviación estándar de la población. La inexactitud de los resultados anteriores se debe a las aproximaciones tomadas.

El proceso de construcción de las tres distribuciones muestrales de medias, asociadas con la población P, y las observaciones hechas al respecto **no** constituyen, en manera alguna, una demostración del resultado que se enunciará a continuación; con lo realizado sólo se intenta hacer evidente, verificar el resultado. Debe tenerse presente que la situación aquí trabajada para una población de cuatro elementos se puede generalizar para cualquier población mucho más grande.

Ahora sí tenemos todos los elementos necesarios para enunciar uno de los resultados mas útiles en estadística y que se conoce como *teorema del límite central*. Dice así:

Considere una población cuya media es μ y cuya desviación estándar es σ. Si de esa población se extraen, al azar, todas las muestras de tamaño n, obtenidas con sustitución y con orden, se puede construir una distribución de medias muestrales, la cual tiene forma aproximadamente normal cuando **n es suficientemente grande**. Además, la media $\mu_{\bar{x}}$ y la desviación estándar EE $(\bar{x})$ de esa distribución muestral están relacionadas con la media y la desviación estándar de la población así:

$$\mu = \mu_{\bar{x}} \qquad \sigma = \sqrt{n}\,EE(\bar{x}) \Rightarrow EE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Nota: La media de la distribución muestral de medias se simbolizará de ahora en adelante con: $\mu_{\bar{x}}$ y la desviación estándar se llamará *error estándar* y se denotará con: EE $(\bar{x})$.

En la práctica, dado que es muy difícil, si no imposible, conocer el valor real de σ con el cual se calcula el valor de EE $(\bar{x})$, suele aproximarse su valor usando la desviación estándar de la muestra con la que se está trabajando.

a. En general, a pesar de que no es posible construir la distribución muestral de medias, sí es posible describirla aproximadamente. Suponga que se tomó una muestra de tamaño 64 de una población y se encontró que la media de la muestra es 538. La desviación estándar de la población es 38. A partir de esa información describa de la manera más completa posible la correspondiente distribución muestral de medias. Ilustre su respuesta. ¿Qué tan probable es que la media de la muestra esté muy alejada de la media de la población? Explique su respuesta.

b. Suponga que se tomó una muestra de tamaño 64 de una población y se encontró que la media de la muestra es 538, y la desviación estándar de la muestra es 43. A partir de esa información describa de la manera más completa posible la correspondiente distribución muestral de medias. Ilustre su respuesta. ¿Qué tan probable es que la media de la muestra esté muy alejada de la media de la población? Explique su respuesta.

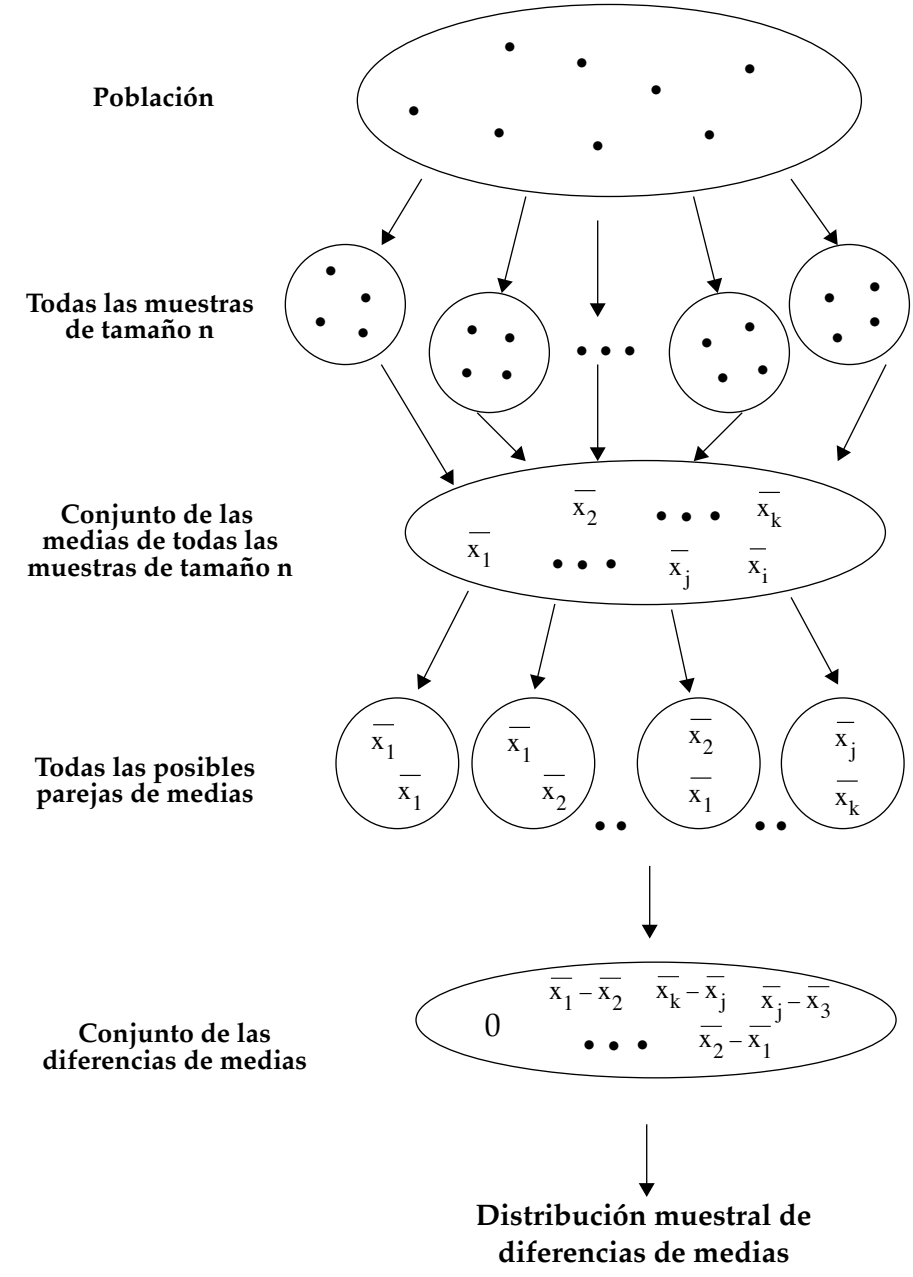
c. En resumen, ¿qué es lo que se requiere saber de la distribución muestral de medias para inferir sobre μ ?

Distribución muestral de diferencias de medias

Este capítulo concluye con la presentación de otro tipo de distribución muestral: la de diferencias de medias. También estas distribuciones son teóricas; en situaciones reales en las que sea necesario comparar dos poblaciones no es posible construir las correspondientes distribuciones muestrales de diferencias de medias, tan sólo es posible imaginarlas. Sin embargo, puesto que el valor a partir del cual se va a inferir sobre las poblaciones pertenece a una de tales distribuciones, se requiere conocer el comportamiento de ellas y además establecer su relación con la distribución de la población. Más concretamente, el manejo de las características de la distribución muestral de diferencias de medias se necesita para determinar si dos muestras provienen o no de la misma población.

Para construir conceptualmente la distribución de diferencias de medias de muestras de tamaño n, se procede así: se extraen, con orden y con sustitución, todas las muestras de tamaño n de la población; se calculan las medias de cada una; se hacen todos los posibles pares de medias (con orden y con sustitución) y se restan. Los valores de diferencia hallados determinan la distribución. El esquema de la página siguiente resume el proceso que se ha expuesto.

A partir de un ejemplo hipotético en el que se conoce la población se construirán dos distribuciones muestrales de diferencias de medias y se destacarán sobre ellas las principales características de ese tipo de distribuciones. El proceso no constituye de ninguna manera una prueba formal. Lo que se busca es verificar, sobre el ejemplo, el enunciado del teorema del límite central. Veamos: la población está constituida por los números 0, 1 y 2, que en este caso se refieren a la temperatura a la que muere un cierto tipo de bacteria. Esta población tiene media 1, desviación estándar 0,8164965 y se distribuye uniformemente en el intervalo [0,2].


Población
Todas las muestras de tamaño n
Conjunto de las medias de todas las muestras de tamaño n
Todas las posibles parejas de medias
Conjunto de las diferencias de medias
0
Distribución muestral de diferencias de medias

A continuación vamos a descubrir cómo se comportan las diferencias de medias de dos muestras del mismo tamaño. En primer lugar, trabajaremos con las muestras de tamaño 2. Al extraer todas las posibles muestras de tamaño 2 con sustitución y con orden se obtienen nueve muestras diferentes. La distribución muestral de medias en este caso se presenta en la siguiente tabla:

Distribución de medias (muestras, n = 2)	
Media	**Frecuencia**
0	1
0,5	2
1	3
1,5	2
2	1

Esta distribución tiene media 1 y desviación estándar 0,5773502. Para construir la distribución muestral de diferencias de medias se van a tomar todos los pares posibles de medias (con orden y con sustitución) y luego se van a restar. Como hay nueve valores de medias, no todos diferentes, se tendrán 81 pares de medias. En el interior de la siguiente tabla de doble entrada se presentan los valores de diferencia obtenidos:

Diferencias entre las medias de las muestras de tamaño 2									
	0	0,5	0,5	1	1	1	1,5	1,5	2
0	0	-0,5	-0,5	-1	-1	-1	-1,5	-1,5	-2
0,5	0,5	0	0	-0,5	-0,5	-0,5	-1	-1	-1,5
0,5	0,5	0	0	-0,5	-05	-0,5	-1	-1	-1,5
1	1	0,5	0,5	0	0	0	-0,5	-0,5	-1
1	1	0,5	0,5	0	0	0	-0,5	-0,5	-1
1	1	0,5	0,5	0	0	0	-0,5	-0,5	-1
1,5	1,5	1	1	0,5	0,5	0,5	0	0	-0,5
1,5	1,5	1	1	0,5	0,5	0,5	0	0	-0,5
2	2	1,5	1,5	1	1	1	0,5	0,5	0

A partir de la tabla anterior se define la distribución de diferencias de medias de muestras de tamaño 2:

Distribución de diferencias de medias de muestras (n = 2)	
Valor de la diferencia de medias	**Frecuencia**
-2	1
-1,5	4
-1	10
-0,5	16
0	19
0,5	16
1	10
1,5	4
2	1

a. Sin hacer cálculos, ¿cuál cree que debe ser la media de esta distribución? ¿Cómo cree que se relaciona la desviación de esta distribución con la de la población? Y, ¿con la de la distribución muestral de medias? Explique el por qué de su intuición.

b. Verifique que la media de la distribución de diferencias de medias para muestras de tamaño 2 es 0 y que la desviación estándar es 0,816496. Represente gráficamente la distribución.

c. Para esa distribución, verifique que en el intervalo de tamaño 1 desviación estándar alrededor de 0 hay 62,96% del total de observaciones y que a 1,96 desviaciones estándar alrededor de la media hay 97,53% del total de observaciones.

Ahora vamos a construir la distribución de diferencias de medias para las muestras de tamaño 3. En este caso hay 27 muestras diferentes, tomadas con

sustitución y con orden. Al calcular la media para cada una de las muestras se obtienen 27 valores, no todos diferentes, a saber: 0, 1/3, 2/3, 1, 4/3, 5/3, 2. La media de esta distribución es 1 y su desviación estándar es 0,4714045. Para obtener la distribución de diferencias de medias se requiere hacer todos los posibles pares de medias y efectuar la resta entre las dos medias. Se tienen para el caso 27* 27 = 729 parejas de medias, a partir de las cuales se obtendrán 729 valores de diferencias de medias, no todos diferentes. A continuación se presenta la distribución de diferencias de medias.

Distribución de diferencias de medias (muestras, n = 3)	
Valor de la diferencia de medias	**Frecuencia**
-2	1
-5/3	6
-4/3	21
-1	50
-2/3	90
-1/3	126
0	141
1/3	126
2/3	90
1	50
4/3	21
5/3	6
2	1

d. Verifique que la media de la distribución es 0 y que la desviación estándar es 0,6666. Además, represente gráficamente la distribución.

e. Para esta distribución, verifique que en el intervalo de tamaño 1 desviación estándar alrededor de la media hay 78,60% del total de los valores de diferencia, y que a 1,96 desviaciones estándar alrededor de la media hay 92,31% del total de observaciones.

En resumen, se han obtenido las dos distribuciones de diferencias de medias para muestras de tamaño 2 y 3, asociadas con la población P. Las características de la población, de las distribuciones muestrales de medias y de las distribuciones muestrales de diferencias de medias se exponen a continuación:

	Distribución	tamaño	media	desv. est.	forma gráfica
	Población	3	1	0,8164965	uniforme
muestras, n = 2	Muestral de medias	9	1	0,5773502	aprox. normal
	Muestral de diferencias de medias	81	0	0,8164965	aprox. normal
muestras, n = 3	Muestral de medias	27	1	0,4714045	aprox. normal
	Muestral de diferencias de medias	729	0	0,666666	aprox. normal

Falta entonces establecer la relación existente entre la desviación estándar de la distribución muestral de diferencias de medias y la desviación estándar de las distribuciones muestrales de medias. Tal relación se puede expresar así:

$$EE(\bar{x} - \bar{y}) = \sqrt{(EE(\bar{x}))^2 + (EE(\bar{y}))^2}$$

donde:

- $EE(\bar{x} - \bar{y})$ representa la desviación estándar de la distribución muestral de diferencias de medias y se le llama el *error estándar de la diferencia o el error estándar diferencial.*
- $EE(\bar{x})$ y $EE(\bar{y})$ representan el error estándar de las correspondientes distribuciones muestrales de medias, siendo que las muestras tienen tamaños similares.

Aunque no constituye ninguna prueba formal, queremos emplear el ejemplo hipotético que hemos venido trabajando para verificar la relación enunciada. Bajo el supuesto de que dos muestras provienen de la misma población P y tienen ambas tamaño 3, se sabe entonces que la diferencia entre las correspondientes medias es un valor que pertenece a la distribución muestral de diferencias de medias, cuya desviación estándar (error estándar de la media) es 0,4714045. Por tanto,

$$EE(\bar{x}-\bar{y}) = \sqrt{(EE(\bar{x}))^2 + (EE(\bar{y}))^2}$$

$$= \sqrt{(0,4714045)^2 + (0,4714045)^2} = 0,666666$$

Y, ése efectivamente es el valor del error estándar diferencial.

En la práctica, dado que es muy difícil, si no imposible, conocer el valor real de σ con el cual se calcula el valor de $EE(\bar{x}-\bar{y})$, suele aproximarse su valor partiendo de la desviación estándar de las muestras usadas.

Aunque los dos ejemplos anteriores no son concluyentes, sirven para cerrar los puntos importantes relacionados con esta distribución:

> Dada una población cuya media es μ de la cual se extraen, al azar, todas las muestras de tamaño n[29], se puede construir una distribución con los valores de diferencia entre las medias de todos los pares de muestras que se pueden tomar. Esta distribución tiene media 0 y es aproximadamente normal en caso de que el tamaño de las muestras sea suficientemente grande. Su desviación estándar se puede estimar a partir del error estándar de dos de las muestras como: $\sqrt{(EE(\bar{x}))^2 + (EE(\bar{y}))^2}$ y se conoce como *error estándar diferencial*.

Es necesario hacer un comentario final con respecto a los errores estándar. Cuando se toman muestras de pocos elementos de una población, sucede que las desviaciones estándar presentan un comportamiento, digamos, errático, en el sentido de que dan valores extremos. Piense qué sucede si en vez de tomar muestras de pocos elementos se toman muestras de más de 30 elementos, por ejemplo. En este caso la desviación que presentan los datos con respecto a la media de la población tiende a ser menor que cuando se toman pocos datos en la muestra. Esto tiene incidencia en el cálculo de los errores estándar de las distribuciones muestrales presentadas en las dos últimas secciones. A mayor número de elementos en las muestras, mejor es la estimación de la desviación estándar de la población.

29 Obtenidas con sustitución y con orden.

Intervalos de confianza

Introducción

En este capítulo se entra en detalles con respecto a una de las tareas que tiene la estadística inferencial: la estimación de parámetros. Inicialmente se presenta y se resuelve un problema en el que se pide la estimación de la media de una población; a partir de su solución se habla de manera informal acerca de los conceptos importantes relacionados con la estimación. Después se formalizan tales conceptos. En la siguiente sección se plantean y se solucionan dos problemas, uno para estimar la media de una población y otro para estimar la diferencia de dos medias. Finalmente, se incluye una sección de ejercicios.

Motivación

Problema. Se extrae,[30] aleatoriamente, de la población P = {1, 2, 3, 5} una muestra de tamaño 4. Suponga que la muestra es M_1 = {2, 1, 2, 1}. Con base en la información que da la muestra estime la media de la población.[31]

a. Imagínese que usted tiene que resolver el problema planteado anteriormente, ¿qué procedimiento seguiría para darle solución? Explique su respuesta. ♣*Para responder esta pregunta piense, por ejemplo, en el significado de las frases "el precio de tal artículo está alrededor de los $1.000"; "espérame entre las siete y las siete y media de la noche".*♣

b. ¿Serviría el procedimiento que usted mencionó en el item anterior para estimar la media de la población con base en la muestra M_2 = {2, 5, 5, 1}?

30 Con orden y con sustitución.

31 Para que el problema tenga sentido vamos a suponer que no conocemos ni la media ni la desviación estándar de la población.

c. Suponga que alguien responde a la pregunta **a.** así:

> "Calculo la media de los valores de la muestra y esa es la estimación que hago de la media de la población."

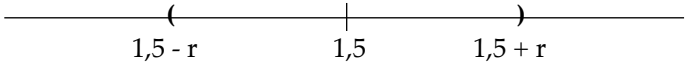
Dé un argumento que muestre que esa no es la mejor solución al problema.

La media de la muestra M_1 es $\overline{x_1} = 1,5$. Es con base en ese valor que se va a realizar la estimación de la media de la población; sin embargo, la estimación que se haga no puede ser, de ninguna manera, una afirmación tajante, categórica; más bien, debe ser una afirmación que dé la idea de entre qué valores es probable encontrar la media de la población. Puesto que la estimación se hace con base en una muestra aleatoria y de muestra a muestra hay variaciones es obvio pensar que la media de la muestra que se haya tomado **no necesariamente** es igual a la media de la población.

De manera que estimar la media de la población consistirá en construir un intervalo cuyo centro sea la media de la muestra que se tiene y cuyo radio sea un determinado valor, r. Para nuestro caso, el intervalo ha de ser:

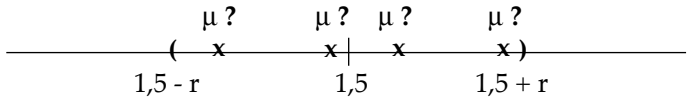
$$[\bar{x} - r; \bar{x} + r] = [1,5 - r; 1,5 + r]$$

Gráficamente se verá así:



Y la interpretación será: se estima que la media de la población se encuentra entre los valores 1,5 - r, y, 1,5 + r.

Con frecuencia se hacen afirmaciones del estilo: "la media de la población varía entre los valores a y b". El valor de μ no varía; la media de la población es un cierto valor fijo, que usualmente se desconoce. En la realidad, lo que varía es la estimación que hacemos de μ. Al establecer un intervalo con centro en la media de la muestra, para estimar la media de la población, se están dando muchísimos valores cercanos a 1,5, en este caso, y se está afirmando que cualquiera de ellos podría ser la media de la población; pero, cuál de ellos es, no se sabe y no se podrá determinar.



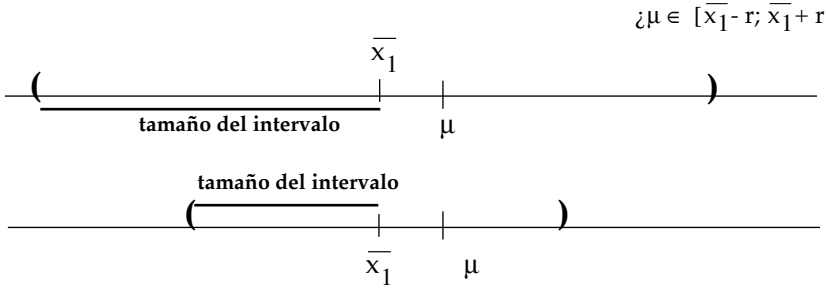
El problema que surge entonces, es ¿cuál debe ser el valor del radio, r, del intervalo? Antes de responder esta pregunta hagamos unas consideraciones:

d. Con respecto a la estimación de la media de una población, comente la siguiente frase, teniendo en cuenta aspectos de precisión y de certidumbre. ♣*Para responder esta pregunta piense en frases como "me demoraré entre 15 y 25 minutos", "me demoraré entre 5 y 35 minutos".*♣

No hay un único intervalo para realizar dicha estimación.

e. ¿Es clara para usted la diferencia entre precisión de un resultado y la certidumbre del mismo? ¿Qué relación hay entre los dos conceptos?

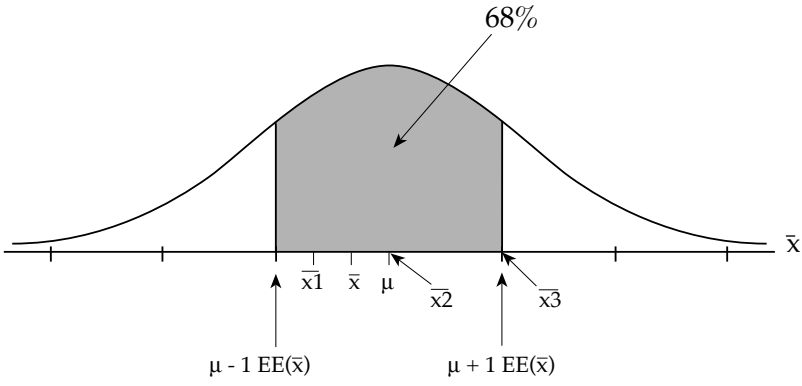
Tal como seguramente usted lo expresó en su respuesta, no existe un único intervalo para estimar la media de la población. Se pueden construir muchísimos intervalos, todos ellos centrados en la media de la muestra y de diferentes tamaños, es decir, de radios diferentes. Algunos de los intervalos incluirán más probablemente que otros, la media de la población. Otros intervalos darán una estimación más precisa de μ. En total, la *certidumbre* y la *precisión* de la estimación son dos conceptos diferentes, pero relacionados entre sí. Entre más precisa sea una estimación, existe menos certidumbre de que sea buena, y recíprocamente, entre menos precisa sea una estimación, existe más certidumbre de que sea buena. En el siguiente esquema se presentan dos estimaciones de μ, hechas a partir del valor $\overline{x_1}$ de una cierta muestra. La primera es menos precisa que la segunda, pero tiene mayor grado de certidumbre.



De todas maneras, para efectos prácticos, es importante tener en cuenta la relación entre esos dos conceptos, y la necesidad de establecer un justo medio que permita perder un poco de certidumbre en aras de ganar en precisión. Es posible que un investigador que quiera tener 100% de confianza en su estimación obtenga un intervalo tan amplio que no le da ninguna información que él no supiera de antemano; en tal caso, si quiere mejorar la precisión de su estimación tendrá que conformarse con un grado de certidumbre menor que el del 100%. Usualmente, los investigadores hacen estimaciones con 90%, 95% y 99% de certidumbre, y eligen un tamaño de muestra que les permita satisfacer un grado de precisión requerido.

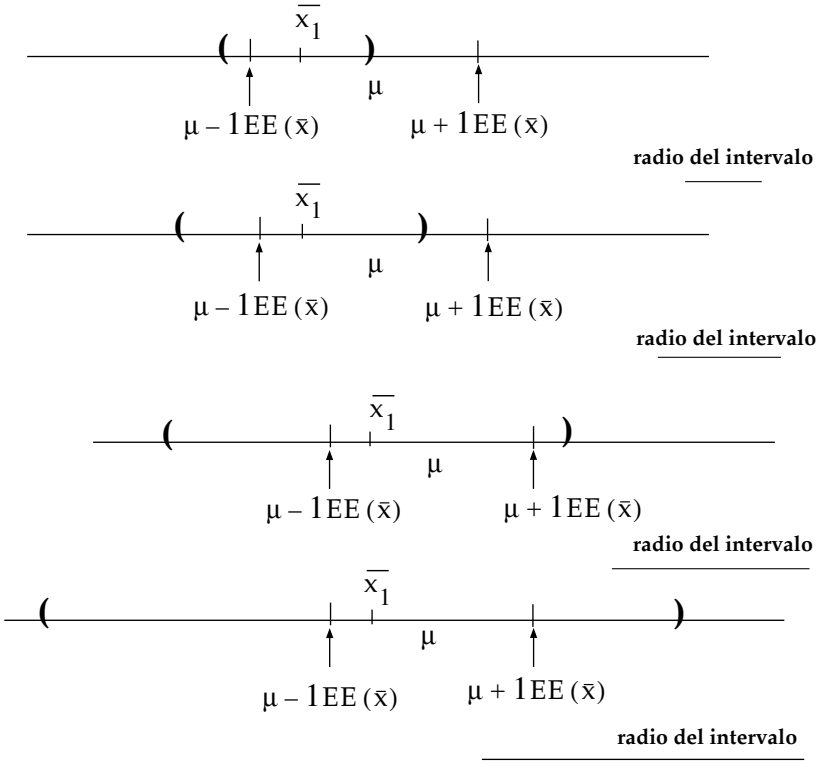
Para la explicación que se presenta a continuación se emplean conceptos y resultados del capítulo anterior, a saber: el valor de la media de la muestra pertenece a una distribución muestral que sigue el modelo normal para casos en que el tamaño de la muestra sea suficientemente grande; además, la media de esa distribución y la de la población son iguales.

En la gráfica siguiente se expone el modelo para la distribución muestral de tamaño 4 y se han señalado cuatro medias muéstrales ($\overline{x_1}$, $\overline{x_2}$, $\overline{x_3}$ y $\bar{x}$). Ninguna de las cuatro medias se aleja de μ más de 1 error estándar (porque así se tomó el ejemplo; sin embargo, podrían haberse tomado diferentes medias muéstrales que estuvieran todas a menos de 1,5 errores estándar de μ, o que todas distaran menos de 2,3 errores estándar de la media, etc.) Además, observe que $\overline{x_2}$ y $\overline{x_3}$ toman posiciones "extremas" dentro del intervalo y $\bar{x}$ representa *cualquier* media muestral del intervalo $[\mu\text{-}1EE(\bar{x}), \mu+1EE(\bar{x})]$.

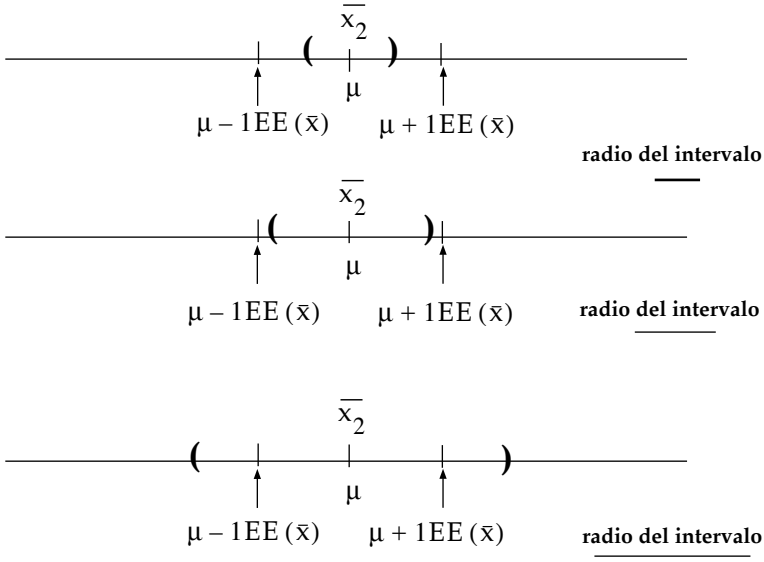


Modelo para la distribución de medias muestrales

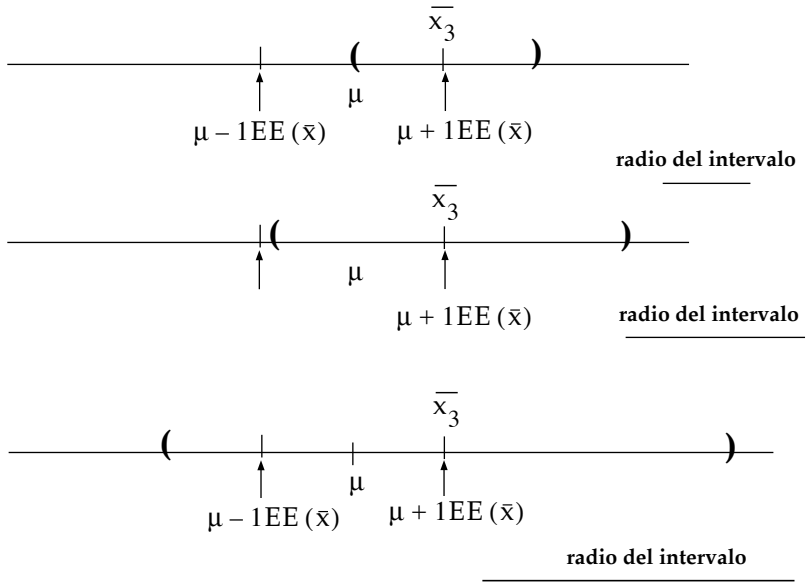
Centremos nuestra atención en $\bar{x}_1$. Vamos a construir un intervalo con centro en $\bar{x}_1$ y tal que incluya a la media de la distribución. Para que eso ocurra, ¿qué radio puede tener el intervalo? Ya sabemos que la respuesta no es única: el radio podría ser igual a la distancia que hay de $\bar{x}_1$ a la media de la distribución. O, podría ser cualquier número mayor que la distancia mencionada. El esquema siguiente muestra cuatro intervalos, cada uno de los cuales tiene centro en $\bar{x}_1$ y su radio es tal que la media de la población, m, pertenece al intervalo. El primer intervalo del esquema es el más pequeño intervalo con centro en $\bar{x}_1$, que contiene a μ.



Para la media $\overline{x_2}$, se tiene una situación especial: por coincidir con la media de la distribución, cualquier intervalo con centro en $\overline{x_2}$ incluirá la media de la distribución, por tanto, el radio es cualquier número positivo. El esquema siguiente muestra tres intervalos, cada uno de los cuales tiene centro en $\overline{x_2}$ y a cada uno de los cuales pertenece μ, sin importar el valor del correspondiente radio.



También para la media $\overline{x_3}$ se tiene una situación especial: si se quiere construir un intervalo centrado en $\overline{x_3}$ que incluya a la media de la distribución, el radio debe tener como valor mínimo la distancia que hay entre tal media y la media de la distribución muestral que para el caso es 1 error estándar. El siguiente esquema muestra tres intervalos, cada uno de los cuales tiene centro en $\overline{x_3}$ y su radio es tal que la media de la población, μ, pertenece al intervalo. El primer intervalo del esquema es el más pequeño intervalo que cumple las condiciones. El radio de los otros intervalos es mayor que 1 EE($\bar{x}$).

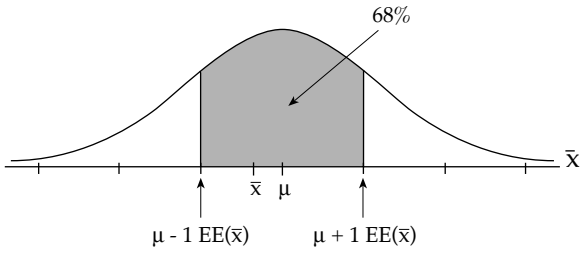


Hemos presentado los valores de tres medias muestrales que están a no más de 1 error estándar y se ha determinado el valor mínimo que debe tener el radio, en cada caso de manera que el intervalo centrado en la correspondiente media incluya la media de la distribución. La respuesta en cada caso ha sido la misma: la distancia que hay entre las dos medias, la de la muestra y la de la distribución muestral. Sin embargo, aunque la respuesta es la misma, el valor no es el mismo. Si quisiéramos dar el mismo valor para el radio de los intervalos centrados en cualquiera de tales medias, el valor mínimo que sirve es 1 error estándar. ¿Por qué?

Entonces, podemos generalizar la situación así: si $\bar{x}$ es una media muestral que dista de la media de la distribución no más de 1 error estándar, entonces se pueden construir infinidad de intervalos con centro en $\bar{x}$. Pero, sólo nos interesan aquellos que contengan la media de la población. Para asegurar que el intervalo que se construye tiene a la media de la población debemos escoger adecuadamente el radio de dicho intervalo. El valor mínimo que debe tener el radio es entonces 1 error estándar.

Se tiene pues que si la media muestral conocida pertenece al intervalo $[\mu\text{-}1EE(\bar{x}), \mu\text{+}1EE(\bar{x})]$ se puede estimar la media de la población, construyendo un intervalo que tenga como centro el valor de dicha media $(\bar{x})$ y su tamaño sea 1 error estándar. Como el modelo que se está empleando es el normal, entonces se sabe que 68% de todas las posibles medias muestrales están en el intervalo $[\mu\text{-}1EE(\bar{x}), \mu\text{+}1EE(\bar{x})]$. Por tanto, se afirma que:

> el intervalo construido $[\bar{x}\text{-}1EE(\bar{x}), \bar{x}\text{+}1EE(\bar{x})]$ estima la media de la población con una probabilidad de acertar del 68%.



Distribución de medias muestrales

La situación presentada con las medias $\overline{x_1}$, $\overline{x_2}$, $\overline{x_3}$ y $\bar{x}$ se puede transferir a medias muestrales que disten más de 1 error estándar de la media de la distribución muestral, logrando así intervalos más amplios, a la vez más confiables, pero menos precisos.

Por ejemplo, si la distancia de la media muestral que se tiene, $\bar{x}$, a la media de la distribución es menos 1,96 errores estándar, entonces un intervalo con centro en tal media muestral y radio igual a 1,96 errores estándar contendrá a la media de la distribución muestral –que es la misma media poblacional– con una probabilidad de acertar del 95%. Y, entonces se dirá:

> se estima que μ está en el intervalo $[\bar{x}\text{-}1{,}96EE(\bar{x}), \bar{x}\text{+}1{,}96EE(\bar{x})]$ con una probabilidad de acertar del 95%.

Una vez tratado lo concerniente al tamaño del radio de los intervalos, podemos dar respuesta al problema propuesto al inicio de esta sección. Reto-

mémoslo: Estime la media de la población de la cual proviene la muestra $M_1 = \{2, 1, 2, 1\}$.

En primer lugar vamos a calcular la media y la desviación estándar de la muestra:

$$\bar{x} = 1{,}5 \text{ y } s = 0{,}5$$

Para estimar el valor de μ vamos a construir un intervalo que tenga centro en 1,5. Vamos a suponer que se quiere tener 95% de certidumbre de que el intervalo así construido contiene a la media de la población. Dado ese hecho, entonces el radio del intervalo debe tener un tamaño igual a 1,96 errores estándar. De manera que el intervalo es:

$$[1{,}5 - 1{,}96\ EE(\bar{x});\ 1{,}5 + 1{,}96\ EE(\bar{x})]$$

Falta buscar el valor del error estándar. El teorema del límite central dice que

$$\sigma = \sqrt{n}\, EE(\bar{x}) \Rightarrow EE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Según esa igualdad, para calcular el error estándar de la distribución muestral de medias se requiere conocer la desviación estándar de la población σ. En este punto se tiene una dificultad pues en la realidad no se conoce la población y por tanto se desconoce el valor de σ. Vamos entonces a obtener un valor aproximado del error estándar, empleando para ello el valor de la desviación estándar de la muestra en vez del valor de la desviación estándar de la población. El tener que estimar el valor de σ con el valor de s, no cambia esencialmente las cosas con respecto a las características de la distribución muestral de medias en caso de que el tamaño de las muestras sea suficientemente grande. Sin embargo, si las muestras son pequeñas y no se conoce el valor de la desviación estándar de la población, entonces la distribución muestral de medias **no** sigue el modelo normal; se distribuye según otro modelo. En este texto nos limitaremos a inferir con base en muestras suficientemente grandes.

Se tiene entonces que:

$$EE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \frac{0{,}5}{2} = 0.25$$

De manera que el intervalo queda:

$$[1,5 - 1,96 * 0,25; 1,5 + 1,96 * 0,25] = [1,01; 1,99]$$

Y se dice que se estima que la media de la población está en el intervalo [1,01; 1,99] con una probabilidad de acertar del 95%. Es importante observar que siempre es posible que el intervalo construido no incluya a μ. Lo dicho se puede observar en el ejemplo que se está trabajando: recuerde que la media de la población es 2,75, valor que no pertenece al intervalo [1,01; 1,99].

f. Compruebe que el intervalo con 99% de probabilidad de acertar, en este caso, tampoco incluye la media poblacional. ¿Qué conclusión cree que se puede obtener del hecho observado?

Ahora considere que la muestra que se tiene es $M_2 = \{2, 5, 5, 1\}$. Vamos a estimar la media de la población con base en ella.

En primer lugar, calculamos la media y la desviación estándar de la muestra:

$$\bar{x} = 3,25 \text{ y } s = 1,785$$

También estimamos el error estándar de la distribución muestral a partir del valor de la desviación estándar de la muestra.

$$EE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{1,785}{\sqrt{4}} = 0,8925$$

Construyamos el intervalo del 95% de certidumbre:

$$[3,25 - 1,96 * 0,8925; 3,25 + 1,96 * 0,8925] = [1,5007; 4,9993]$$

Por tanto, se estima, con 95% de confianza de acertar, que la media de la población está en el intervalo [1,5007; 4,9993]. Y observe que en este caso, el intervalo construido **sí** incluye a la media de la población.

Formalización de algunos conceptos

En la sección anterior se hizo mención siempre a la estimación de la media de una población y debió quedar claro el proceso de construcción de intervalos a través de los cuales se hace la estimación. Pues bien, la esencia de ese proceso es la misma siempre que se quiera estimar un parámetro. Sólo cambian el modelo que se emplea (normal, u otros que aquí ni siquiera nombraremos) de acuerdo al comportamiento de la distribución muestral correspondiente, y el error estándar. En resumen, ese proceso se puede expresar así:

> estimador ± r, donde r depende de dos aspectos de la distribución muestral implicada en el caso (modelo y fórmula para calcular el error estándar) y también depende del grado de certidumbre que se quiera tener.

Por tanto, aunque no se haya dicho nada explícito sobre la estimación de la diferencia de dos medias, el proceso no tiene nada nuevo con respecto al que mostramos para la estimación de la media poblacional. Para el nuevo caso, se trabajará con la distribución muestral de diferencias de medias, la cual sigue el modelo normal en caso de que las muestras sean suficientemente grandes y tiene error estándar dado por la fórmula $\sqrt{(EE(\bar{x}))^2 + (EE(\bar{y}))^2}$. Además, las muestras deben tener tamaños similares.

En la próxima sección se dará un ejemplo de estimación de la diferencia de dos medias. En esta sección, que será muy breve, trataremos de definir de manera general tres conceptos relativos a la estimación de parámetros.

> Un *intervalo de confianza* para estimar un parámetro, a partir del correspondiente valor del estadístico de una muestra (estimador), es un intervalo con centro en dicho estimador y un cierto radio. El intervalo, probablemente incluye el parámetro de interés.
> El *error* o *precisión* del intervalo es el tamaño de su radio.
> El *nivel de confianza* del intervalo es la probabilidad que se tiene de acertar en la estimación. Un nivel de confianza del x% significa que en x% de las ocasiones en las que se construye el intervalo de confianza, el parámetro estará incluido ahí.

Dos ejemplos

Problema 1. Para determinar la rentabilidad de un nuevo restaurante, un investigador observó durante 30 días, las ganancias del mismo. Encontró que la ganancia media era de $20.000 diarios con una desviación estándar de $3.000 diarios. ¿Cuál es la ganancia diaria promedio, con un nivel de confianza del 90%?

Tamaño de la muestra	$n = 30$
media de la muestra	$\bar{x} = 20.000$
desviación estándar de la muestra	$s = 3.000$
error estándar de la distribución muestral	$EE(\bar{x}) \approx \frac{3.000}{\sqrt{30}} = 547,44$

Puesto que se quiere un nivel de confianza del 90% y el modelo de la distribución de medias muestrales es normal, se sabe que el valor del puntaje z es de 1,64. Por tanto, el intervalo de confianza es:

$$[20.000 - 1,64 * 547,44; 20.000 + 1,64 * 547,44] = [19.102,20; 20.897,80]$$

De manera que se estima, con un nivel de confianza del 90%, que la ganancia diaria promedio del restaurante está entre $19.102,20 y $20.897,80.

Problema 2. En un centro de estética, durante los últimos seis meses, se han estado empleando dos tratamientos diferentes para reducir de peso (T1 y T2). El tratamiento T1 se ha aplicado a un grupo G1, mientras que el tratamiento T2 se ha aplicado a un grupo G2. Ambos grupos están formados por adultos cuyas edades oscilan entre 25 y 35 años, que tienen problemas de obesidad. El tratamiento T1 es sustancialmente más costoso que el T2. El médico del centro quiere determinar entre qué par de valores se puede esperar que esté la diferencia en los pesos medios rebajados después de los tratamientos para tomar decisiones hacia el futuro con respecto al tratamiento que debe ofrecer el centro. Al terminar la aplicación de los tratamientos, los resultados obtenidos son los siguientes:

	Muestra 1	**Muestra 2**
Tamaño	n = 49	n = 49
Media	$\bar{x}$ = 16,8 kilos	$\bar{y}$ = 15,24 kilos
Varianza	s = 3,5 kilos	s = 3,3 kilos

El proceso para resolver este problema es similar al anterior. Sin embargo, como lo que ahora se quiere estimar es la diferencia de medias entonces la estimación se basará en el comportamiento de la distribución muestral de diferencias de medias. Puesto que esta distribución es aproximadamente normal, entonces también emplearemos el modelo de la normal. Además, vamos a usar un nivel de confianza del 96%. Veamos la solución:

Diferencia de medias, $\bar{x} - \bar{y} = 1{,}56$ kilos

Error estándar de la distribución muestral de diferencias de medias,

$$EE(\bar{x} - \bar{y}) = \sqrt{\frac{10,89}{49} + \frac{12,25}{49}} = 0,6872$$

Puesto que se quiere un nivel de confianza del 96% y el modelo de la distribución muestral de diferencias de medias es normal, se sabe que el valor del puntaje z es de 2,05. Por tanto, el intervalo de confianza es:

$$[1{,}56 - 2{,}05 * 0{,}6872;\ 1{,}56 + 2{,}05 * 0{,}6872] = [0{,}1512;\ 2{,}9687]$$

De manera que se estima, con un nivel de confianza del 96%, que la diferencia en los pesos medios rebajados después de los tratamientos varía entre 0,15 kilos y 2,96 kilos. Puesto que este intervalo no incluye a 0, –caso en el que se registraría que no hay diferencia significativa en la efectividad de los tratamientos– entonces, parece ser que uno de los tratamientos –el T1– produce mejores resultados que el otro.

Ejercicios

1.- Se examinó una muestra de 36 cigarrillos de cierta marca para determinar el contenido de nicotina. La muestra tuvo una media de 22 miligramos y una desviación estándar de 4 miligramos. Estime, con 90% de confianza, la media del contenido de nicotina de los cigarrillos de esa marca. También estime con 95% y con 99% la media del contenido de nicotina. Haga comentarios pertinentes sobre los resultados que encuentra con respecto al hecho de aumentar el nivel de confianza.

2.- Se examinó una muestra de 64 cigarrillos de cierta marca para determinar el contenido de nicotina. La muestra tuvo una media de 22 miligramos y una desviación estándar de 4 miligramos. Estime, con 90% de confianza, la media del contenido de nicotina de los cigarrillos de esa marca.

3.- Compare los resultados de los dos problemas anteriores y haga comentarios pertinentes con respecto al hecho de aumentar tamaño de la muestra.

4.- Determinar el intervalo de confianza del 99% para el peso medio de los venusinos, si se sabe que una muestras aleatoria de 36 venusinos arrojó la siguiente información (con respecto al peso de ellos, medido en kilogramos). El peso se distribuye normalmente.

16	22	31	28	15	20	20	21	22	35	28	27
25	24	20	18	19	31	17	18	20	15	25	24
27	18	20	31	24	23	20	20	19	31	30	20

5.- En cierto municipio colombiano un grupo de investigadores sociales busca determinar la cantidad promedio del impuesto al valor agregado en los establecimientos de comida y con ello estimar el ingreso que recibe el municipio por razón del IVA. Para dicho estudio se tomó una muestra de 36 establecimientos (restaurantes) de la población de estudio y se encontró que el ingreso promedio semanal, por razón del IVA, en esa muestra, fue de $65.736 con una desviación estándar igual a $5.402.

a. ¿Cuál es la población y la muestra de estudio?

b. ¿Qué variable se está midiendo y de qué tipo es?

c. ¿Cuál es el objetivo de la investigación?

d. Estime entre qué valores se encuentra el impuesto promedio que recaudó el municipio por concepto del IVA. Haga una estimación con 99% de probabilidad de acertar.

6.- Mundialmente se ha reconocido que aquellos niños que tienen un coeficiente intelectual igual o superior a 125 son superdotados. En el colegio X hay cinco secciones del curso tercero elemental. De ese grupo de alumnos se tomó al azar una muestra de 30 niños y se encontró que el coeficiente intelectual promedio es de 116,5 con una desviación estándar de 14,1.

a. Si se sabe que en la muestra, el coeficiente intelectual se distribuye normalmente, ¿cuántos niños se pueden considerar superdotados?

b. Utilice la muestra tomada para estimar el coeficiente intelectual promedio de los alumnos de tercero elemental de dicho colegio.

c. Sugiera rangos que permitan clasificar a los niños como subnormales, normales y superdotados.

7.- En el siglo XVI se formó el Palenque de San Basilio, comunidad afroamericana de negros cimarrones que aún a finales del siglo XX sigue conservando el vernáculo palenquero, lengua criolla[32] de base léxica puramente española.

Para analizar el proceso de cambio de esta lengua por influencia del español moderno, unos etnolingüistas estudiaron el criollo palenquero de San Basilio durante 1990, recolectando un corpus compuesto por 500 conversaciones cotidianas de los habitantes de la población. Se miró el porcentaje de palabras propias de la lengua empleadas y se encontró que tales porcentajes se distribuían normalmente con una media igual a 63,1% y una desviación estándar de 10,2%.

a. En un estudio realizado diez años atrás se encontró que el porcentaje promedio de uso del palenquero era del 88%. En el lapso de diez años, ¿ha cambiado mucho el lenguaje?

32 Una lengua criolla es aquella que surge por el contacto entre una lengua dominante y una más débil. El palenquero viene de la mezcla entre el español impuesto por los colonizadores y un dialecto africano hablado por una minoría de esclavos negros.

b. Los etnolingüistas esperaban que la media de la población se encontrara entre el 70% y el 75% con una probabilidad del 95%, para afirmar que no ha habido un cambio significativo en el lenguaje. ¿Los datos presentados confirman lo esperado por los etnolingüistas?

8.- En un estudio psicológico sobre la susceptibilidad a las ilusiones perceptivas, 50 hombres juzgan la longitud de una figura ilusoria. La evaluación de cada uno de ellos se compara con la longitud verdadera y se registra la diferencia. El experimento produjo los siguientes resultados para las diferencias: $\bar{y} = 81$ milímetros; $s = 12$ milímetros. Encuentre un intervalo de confianza del 95% para la magnitud media de las diferencias.

9.- Coca-Cola y Postobón compañías en eterna competencia, lanzaron ofensivas propagandísticas para promocionar "agua manantial" y "agua cristal" en la ciudad de Bogotá. Un estudiante que hacía su tesis sobre "Propaganda y Cia. Multinacionales" quiso determinar cuál de las dos campañas había sido más efectiva. Para eso, tomó dos muestras de habitantes de barrios de clase media del norte de la ciudad.

Entrevistó a 500 consumidores de "agua manantial" y halló que el consumo promedio era de 5,3 litros por semana, con una desviación estándar de 1,41 litros. También entrevistó a 400 consumidores de "agua cristal" y halló que el consumo promedio era de 5,6 litros por semana, con una desviación estándar de 1,73 litros.

a. ¿Cuál es la población y cuáles son las muestras de estudio?

b. Con base en la información que dan las dos muestras estime la diferencia media de consumo en las poblaciones de consumidores mencionadas. Utilice un nivel de confianza del 95%.

c. Utilice el intervalo de confianza construido en el item anterior para determinar si se puede concluir que alguna de las dos campañas fue mejor que la otra.

10.- Para comparar dos hospitales en cuanto a la eficiencia en la atención a pacientes, en cada uno de ellos se toman 100 mediciones del número de pacientes atendidos mensualmente. Los resultados obtenidos se consignan a continuación:

Hospital A	
promedio de pacientes atendidos mensualmente	1.000
desviación estándar	100
Hospital B	
promedio de pacientes atendidos mensualmente	1.050
desviación estándar	80

a. Estime la media del número de pacientes atendidos en el hospital A, en un intervalo de confianza del 99,5%.

b. Utilice la información del intervalo de confianza construido en el item anterior para determinar si hay diferencia significativa en cuanto a la eficiencia de los hospitales.

11.- Desde fines del siglo XIX numerosos pedagogos se han interesado por los métodos de enseñanza de lenguas extranjeras. Hoy, que se hace imperante aprender un idioma distinto al nativo, se le da mucha más importancia a la efectividad de tales métodos.

Con el fin de comparar los métodos de dos centros especializados en idiomas, un profesor de lenguas modernas adelantó un estudio que examinaba la calidad del idioma que 39 estudiantes de cada centro habían aprendido. Los exámenes, cuyas calificaciones sobre 70 resultaron distribuirse de forma aproximadamente normal, arrojaron los siguientes resultados:

	Centro 1	**Centro 2**
media, $\bar{X}$	53,2	59,8
desviación estándar, s	5,1	3,5
tamaño de la muestra, n	39	39

a. Hay muchas personas que afirman que el Centro 1 es mejor que el Centro 2. ¿Es esto cierto? Justifique su respuesta.

b. Se esperaba que la diferencia en cuanto a la efectividad de los dos métodos estuviese entre -1,8 y -0,3. ¿Lo encontrado en el estudio confirma esta expectativa?

c. ¿Entre qué par de valores está la diferencia de los métodos?

Pruebas de hipótesis

Introducción

Los dos capítulos anteriores se dedicaron al tema de la estimación de intervalos de confianza. Vimos, por ejemplo, el caso de estimación de un intervalo de confianza para una media poblacional, μ. En este capítulo se hace una introducción al tema de pruebas de hipótesis. Sólo trataremos el caso particular de pruebas de hipótesis para una media poblacional, μ.

La metodología de pruebas de hipótesis está íntimamente relacionada con la de intervalos de confianza; de hecho, como veremos, se puede verificar una hipótesis estadística usando intervalos de confianza. Sin embargo, en problemas donde surge la necesidad de toma de decisiones es más natural utilizar en primera instancia la metodología de pruebas de hipótesis.

A continuación se presenta el caso del juicio contra Tahuro. En este juicio Tahuro es acusado de jugar con una moneda que está cargada. Se parte entonces del supuesto de que Tahuro es inocente y la fiscal (Ana Liza) debe allegar información para demostrar la culpabilidad de nuestro amigo; y la defensa (Stadi Shka) tratará de defender a Tahuro; el profesor jugará un papel neutral en el caso: será el juez.

El juicio contra Tahuro se presenta en tres partes: en la primera se describe la acusación contra el joven; en la segunda, se busca un criterio para juzgarlo; y por último, en la tercera se emite el veredicto. Ahora bien, la idea es emplear la situación planteada para aproximarnos a los conceptos elementales relacionados con las pruebas de hipótesis, es así como después del juicio —las tres partes antes descritas— se formalizan los conceptos involucrados, posteriormente se describe el proceso mismo de la metodología de pruebas de hipótesis, luego se presenta un ejemplo y finalmente se formulan algunos problemas.

Motivación: Juicio contra Tahuro

Vamos a comenzar el estudio de una de las herramientas más conocidas de la inferencia estadística: la *prueba de hipótesis*.[33] La forma en que ésta se utiliza tiene gran similitud con lo que se plantea en un juicio contra un acusado. En los juicios se parte de un principio: **la inocencia del acusado** y se procede de tal manera que un personaje, conocido como el fiscal, intenta recoger información para demostrar la culpabilidad del acusado. Similarmente, en las investigaciones donde se utilizan pruebas de hipótesis, se parte de un supuesto básico: la *hipótesis nula* y el investigador trata entonces de recoger información, con base en una muestra aleatoria, para poder decidir si rechaza o no la mencionada hipótesis. En caso de rechazarla, se acoge a otra hipótesis conocida como la *hipótesis alternativa*.

El principio de inocencia

(*El fiscal dirigió su mirada hacia el Jurado.*)

El fiscal: Las pruebas que recaen en contra del acusado permiten concluir que éste es culpable.

El defensor: Protesto, señores del Jurado. Quiero recordarles que la ley dice que hasta que no se demuestre lo contrario, el acusado debe considerarse como inocente y las pruebas que presenta el fiscal aunque permiten sospechar del acusado, no me parecen suficientes para declararlo culpable.

El fiscal: La defensa alega que faltan pruebas para declarar al acusado como culpable y que por tanto debe considerársele como inocente. Entonces, reuniré más pruebas hasta lograr demostrar que el acusado sí es culpable. Estas pruebas serán tan contundentes que la misma defensa no tendrá más remedio que aceptar la culpabilidad.

El defensor: Entonces, señores del Jurado, esperemos a que el fiscal pueda conseguir esas pruebas, pues de lo contrario, y aunque el acusado fuera realmente culpable, no se le puede declarar culpable.

33 Algunos autores prefieren utilizar la palabra verificación o docimasia en vez de la palabra prueba, pues el sentido en que se emplea esta palabra en estadística es bien diferente de lo que se entiende en matemáticas como prueba o demostración formal.

El juez: Por hoy, se cierra la sesión.

Tahuro es acusado: primera parte del juicio

Profesor: Se inicia la sesión. Señor Tahuro, por favor, póngase de pie. (*Tahuro se levanta del puesto de los acusados*.) Se le acusa a usted de estar jugando por los corredores de la universidad, con una moneda que está cargada. ¿Cómo se declara usted ante esa acusación?

Tahuro: Inocente, señor juez.

Profesor: Tiene usted la palabra señorita fiscal.

Ana Liza: El señor Tahuro nunca va a clase de Probabilidad y Estadística. En vez de ello se pasa la vida jugando con una monedita por los corredores de la universidad y varios testigos pueden confirmar lo que digo. En todo caso, no se le acusa de que falte a clase o de que juegue con su monedita, sino de que deja a sus compañeros sin dinero para almorzar.

Stadi Shka: Protesto su señoría. Sí, es cierto que Tahuro falta con frecuencia a clase, pero no es cierto que ande robando el dinero a sus compañeros. El obtiene por medio de un juego limpio y legal sus ganancias, con las cuales paga parte de su matrícula.

Profesor: Ha lugar, señorita fiscal, explique más detalladamente por qué usted acusa al señor Tahuro de robarle dinero a sus compañeros.

Ana Liza: Su señoría, el señor Tahuro roba a sus compañeros utilizando una moneda que está cargada; mejor dicho, usa una moneda para la cual la probabilidad de obtener "cara" no es igual a la de obtener "sello".

Stadi Shka: Protesto, su señoría. La fiscal acusa a Tahuro de usar una moneda cargada, sin tener pruebas; y hasta que no se demuestre lo contrario, debemos suponer que la moneda no está cargada y que por consiguiente la probabilidad de obtener "cara" o "sello" es igual a $1/2$. ♣*Aquí la defensa está usando el principio de inocencia.*♣

Profesor: Señorita fiscal, ¿tiene usted pruebas de que la moneda está cargada?

Ana Liza: No. Pero, puedo demostrarles que la moneda está cargada, utilizando probabilidades y estadística. Sólo necesito que su Señoría me permita

repetir varias veces un experimento aleatorio consistente en lanzar la moneda y observar el resultado que se obtiene en cada ocasión. ♣*Aquí la fiscal asegura que va a conseguir las pruebas para demostrar la culpabilidad de Tahuro.*♣

Profesor: Señorita fiscal, aquí tiene la moneda. (*El profesor entrega la moneda a Ana Liza*.)

Ana Liza: Señor juez, lanzaré esta moneda cien veces y alguno de ustedes contará el número de veces que se obtiene "cara".

Stadi Shka: Protesto su señoría. El azar puede jugarnos una mala pasada. Por ejemplo, podría ocurrir que por puro azar, todas las veces se obtuviera "cara".

Ana Liza: Lo que alega la defensa es cierto; pero yo les advertí que mi argumento hará uso de la estadística y de la probabilidad. Si la moneda no está cargada y se lanza, por ejemplo, cien veces, la probabilidad de que siempre caiga en "cara" es de:

$$(1/2)^{100} = 0{,}000...0007886$$ ♣*!Huy! Van 30 ceros después de la "coma".*♣

Tal cifra indica que es muy, muy remoto que eso ocurra. Entonces, si se supone la inocencia de Tahuro, es decir, si creemos la hipótesis de que en su moneda, las probabilidades de "cara" y "sello" son iguales a $1/2$ y realizamos el experimento de lanzar cien veces la moneda, la ocurrencia de una proporción exagerada de "sellos" contra "caras" o viceversa da lugar a pensar que la moneda sí está cargada.

Stadi Shka: El argumento del fiscal es muy interesante, pero de ninguna manera nos permite demostrar rotundamente la culpabilidad de Tahuro en caso de que ocurra una proporción exagerada de "sellos" contra "caras" o viceversa, por las siguientes razones:

- Primero, no se ha dicho a partir de qué punto la mayor ocurrencia de "caras" o de "sellos" se considera como exagerada.
- Segundo, si asumimos la hipótesis de que la moneda no está cargada, y que al lanzar la moneda cien veces, siempre se obtiene "cara", tendremos que, aunque es un resultado muy, muy improbable, sí puede ocurrir por un alocado azar.

Profesor: Yo como juez, debo ser imparcial. Realmente pienso que ambos argumentos, tanto el de la defensa como el de la fiscalía son válidos. En todo caso, propongo que la fiscal lleve a cabo su experimento, pero fijando de antemano el criterio que usaremos para considerar como muy exagerado el desequilibrio entre la proporción de "caras" y "sellos".

Stadi Shka: Pero insisto señor juez en que si existe un desequilibrio muy exagerado en la proporción de "caras" y "sellos", podemos en todo caso incurrir en un error que aunque muy poco probable, nos llevaría a la garrafal equivocación de declarar como culpable a una persona que realmente es inocente.

Ana Liza: Señor juez, también existe la posibilidad de que Tahuro sea realmente culpable y que por puro azar no se le declare culpable.

Stadi Shka: ¡Protesto su señoría! La fiscal está tratando de... (*Ana Liza no deja terminar a Stadi Shka.*)

Ana Liza: De ninguna manera. La defensa es la que está tratando de...(*Mientras tanto Stadi Shka continúa hablando, entonces el juez toma su borrador y da varios golpes contra la mesa.*)

Profesor: !Orden en la sala! En este juicio, usando probabilidad y estadística, nunca podremos estar completamente seguros de nuestra decisión: nos podemos equivocar de dos maneras diferentes, a saber:

- Declarar culpable a Tahuro, siendo realmente inocente.
- No declarar culpable a Tahuro, siendo culpable.

En todo caso, vamos a continuar con este juicio y trataremos de determinar a partir de qué punto se va a considerar como exagerado el desequilibrio en las proporciones de "caras" y "sellos".

Stadi Shka: Está bien, estoy de acuerdo su señoría, pero recordemos que tenemos que admitir la posibilidad de obtener cien "caras" en cien lanzamientos de una moneda legal sólo por pura casualidad.

Ana Liza: Pero entonces, ¿pueden ocurrir, por ejemplo, 65 "caras" o más sin que se considere necesariamente que Tahuro es tramposo? En todo caso, debemos reconocer que la probabilidad de que salgan cien "caras" es tan pequeña e insignificante que cualquier persona estaría dispuesta a afirmar con mucha seguridad que este resultado nos indica que la moneda está cargada.

Profesor: Creo que estamos llegando a un acuerdo. Pero, por hoy vamos a cerrar la sesión. Para la próxima sesión queda abierto el problema de:

> Determinar un criterio para establecer a partir de qué punto la moneda de Tahuro puede considerarse como ilegal. Es decir, determinar el número mínimo y el máximo de caras que pueden ocurrir, a partir de los cuales se considerará que la moneda está cargada.

Búsqueda del criterio y final del juicio

Las siguientes preguntas se proponen con la intención de guiar de manera general la reflexión acerca de lo que es el criterio para tomar la decisión en el juicio y las características que debe tener.

a. ¿Cree usted que si en los cien lanzamientos de la moneda se obtienen 53 caras, eso representa evidencia de que la moneda está cargada? Y, ¿si se obtienen 56 caras? Y, ¿si se obtienen 58 caras? Y, ¿si se obtienen 75 caras? Y, ¿si se obtienen 43 caras? Y, ¿si se obtienen 41 caras? Y, ¿si se obtienen 25 caras?

b. Proponga un criterio para decidir sobre la legalidad de la moneda de Tahuro. Explique detalladamente su respuesta (qué razones tiene para proponer ese criterio, por qué lo enuncia así, comentarios adicionales).

c. Considere los siguientes dos criterios para juzgar la legalidad de la moneda de Tahuro:

Criterio 1: Si salen menos de 40 caras, o, si salen más de 60 caras, entonces la moneda de Tahuro está cargada.
Criterio 2: Si salen menos de 35 caras, o, si salen más de 65 caras, entonces la moneda de Tahuro está cargada.

¿Cuál es la diferencia entre los dos criterios? ¿Con cuál de los dos criterios preferirá Tahuro que se le juzque? ¿Por qué?

d. Considere los siguientes dos criterios para juzgar la legalidad de la moneda de Tahuro:

Criterio 2: Si salen menos de 35 caras, o, si salen más de 65 caras, entonces la moneda de Tahuro está cargada.
Criterio 3: Si salen más de 64 caras, entonces la moneda de Tahuro está cargada.

¿Cuál es la diferencia entre los dos criterios? Si la acusación que se ha hecho sobre la moneda de Tahuro es: "la moneda está cargada", ¿cuál de los dos criterios es más adecuado para decidir en el juicio? Explique su respuesta. ¿De qué estilo debería ser la acusación contra la moneda de Tahuro para que el criterio 3 resultara adecuado?

e. Suponga que ya se ha adoptado un criterio para juzgar la legalidad de la moneda de Tahuro. Se lleva a cabo la experiencia que Ana Liza propuso, es decir, se lanza la moneda cien veces y se cuenta el número de caras obtenidas. Además, se aplica el criterio establecido previamente. ¿Qué opinión le merece a usted el hecho de repetir la experiencia, para volver a aplicar el criterio? Explique su posición.

f. A partir de las respuestas dadas a las preguntas anteriores haga una lista de características que debe tener el criterio que se adopte para tomar la decisión en el juicio contra la moneda de Tahuro.

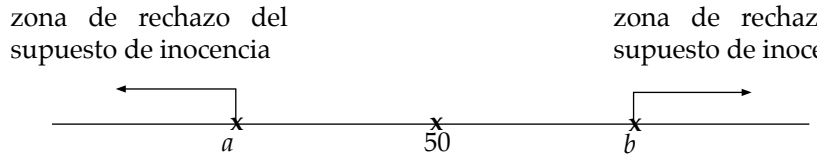
Desde el punto de vista teórico, si una moneda corriente se lanza cien veces, se espera obtener 50 caras y 50 sellos. Sin embargo, si en la práctica se obtuvieran, por ejemplo, 53 caras y 47 sellos, seguramente este hecho no daría pie para sospechar que la moneda está cargada. En este caso, podríamos aceptar la diferencia entre lo teórico y lo experimental como una consecuencia de la presencia del azar en el experimento que se está realizando. De manera similar, no se pensaría que la moneda está cargada si se obtuvieran 47 caras y 53 sellos. Y, podríamos seguir dando casos particulares en los que al lanzar cien veces la moneda no se obtienen 50 caras y 50 sellos y no por eso se sospecha de la legalidad de la moneda. Surge entonces la pregunta: ¿en qué casos, tiene sentido sospechar de la legalidad de la moneda? La respuesta expresada de manera muy vaga sería: en casos en los que el número de caras (y, por tanto, también el número de sellos) sea "muy diferente" de 50.

Antes de llegar a determinar con precisión el criterio que estamos buscando vamos a ponernos de acuerdo en el tipo de criterio adecuado para la situación en la cual se va a emplear. Puesto que la sospecha que se tiene sólo se refiere a que la moneda puede estar cargada y no a que la moneda puede estar cargada a favor de un determinado resultado, entonces ese hecho debe reflejarse en la

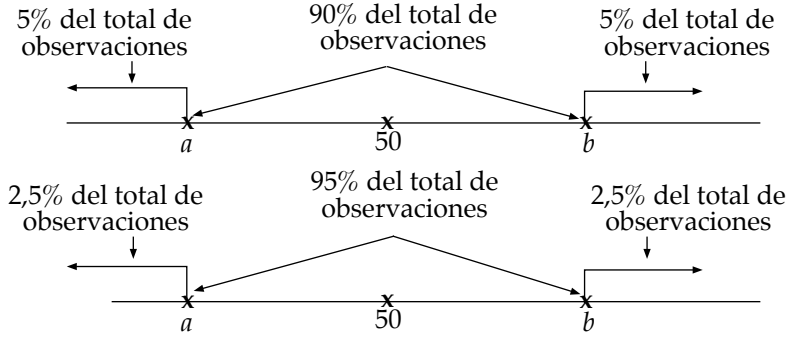
norma que vamos a usar. Así, pues, el criterio deberá expresar que en cualquiera de los siguientes casos se rechazará el supuesto de que la moneda es legal:

- si se obtienen menos de a caras
- si se obtienen más de b caras

siendo que $0 < a < 50$ y $50 < b < 100$



Ahora bien. Los valores a y b no son únicos. Ellos dependen de qué tan rígido se quiere que sea el criterio. Entre más próximos estén a y b a 50, más estricto es el criterio con el que se juzga el resultado y hay más probabilidad de rechazar el supuesto de que la moneda no está cargada; por tanto, hay más probabilidad de acoger la hipótesis de que la moneda sí está cargada, cuando en realidad podría ser una moneda corriente. Recíprocamente, entre más razonablemente alejados estén a y b de 50, el criterio con el que se juzga el resultado de los cien lanzamientos es más amplio, puesto que incluye resultados diferentes a 50, que pueden suceder por azar y no necesariamente porque la moneda esté cargada. La idea es que a y b estén separados de 50 lo necesario para incluir entre ellos una buena proporción del total de los resultados posibles y no incluir una baja proporción de resultados, que son los que se pueden considerar como atípicos. El esquema siguiente puede aclarar lo dicho anteriormente.



A continuación procederemos a precisar el criterio. Para ello vamos a imaginar que el experimento consiste en lanzar una moneda corriente cien veces consecutivas y registrar el números de caras obtenidas. Ese experimento se repite una gran cantidad de veces. La repetición del experimento produce una distribución con las siguientes características:[34]

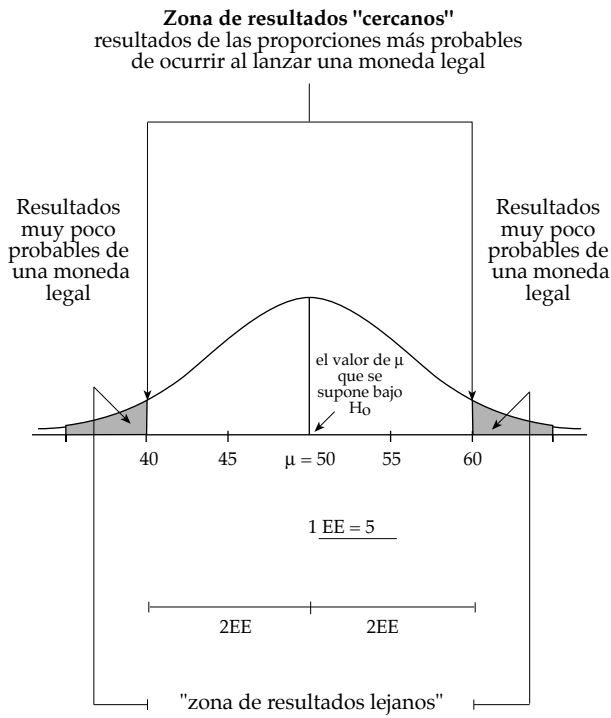
- la variable, el número de caras, toma valores desde 0 hasta 100
- la media de la distribución es 50 caras
- la desviación estándar de la distribución es 5 caras
- la distribución sigue el modelo normal

Y, entonces sabemos por ejemplo que a 2 desviaciones estándar alrededor de la media se encuentra el 95,44% del total de las observaciones. Es decir, en el intervalo [50 - 2 * 5, 50 + 2 * 5] = [40, 60] se incluyen 95,44% de los resultados posibles. El criterio, podría ser el siguiente:

> Si se obtienen menos de 40 caras ó más de 60 caras en cien lanzamientos consecutivos de la moneda, se considera que la moneda es ilegal y por tanto deberá declararse culpable a Tahuro.

La representación gráfica del criterio con el cual se va a juzgar la legalidad de la moneda de Tahuro se presenta a continuación:

34 En el texto no vamos a deducir tales características. La razón es que, aunque no resultaría difícil hacerlo, nos apartaría del tema que se está desarrollando.



Decisión: veredicto

Profesor: Se inicia la sesión. Señor Tahuro, por favor, póngase de pie. (*Tahuro se levanta del puesto de los acusados*.) Se le acusa a usted de estar jugando por los corredores de la universidad, con una moneda que está cargada. ¿Cómo se declara usted ante esa acusación?

Tahuro: Inocente, señor juez.

Profesor: Tiene usted la palabra señorita fiscal.

Ana Liza: Tal como el señor juez lo propuso he determinado un criterio para decidir hasta qué punto una moneda puede considerarse como ilegal. El criterio dice: "Si salen más de 60 caras o menos de 40 caras en cien lanzamientos consecutivos debemos considerar que la moneda es ilegal".

Profesor: El criterio propuesto por la fiscal, coincide con mi criterio. Por tanto, creo que llegó la hora de realizar el experimento.

Stadi Shka: Su señoría, propongo que el experimento sea realizado por un testigo neutral; sugiero que sea Chiripa quien haga los lanzamientos. ♣*El nunca ha sido perjudicado con el asunto de los almuerzos.*♣

Profesor: Se acepta la moción. Señor Chiripa, pase al banquillo de los testigos.

(*Chiripa pasa al banquillo, toma la moneda de Tahuro y realiza los cien lanzamientos, obteniendo 62 "caras".*)

Ana Liza: Señor juez, se han obtenido 62 caras. Este resultado o uno en el que se obtengan más de 60 "caras", sólo ocurre con probabilidad menor de 0,025, por tanto creo que debe declararse a Tahuro como culpable, pues el resultado obtenido no apoya la hipótesis de que la moneda sea normal.

Stadi Shka: ¡Protesto, su señoría! Deberíamos repetir el experimento pues 62 caras se pueden obtener por puro azar aún si la moneda es legal. Entonces me parece que si se vuelve a repetir el experimento, un resultado muy similar al anterior indicaría que...

Ana Liza: No es necesario que siga justificando la defensa. Que se repita el experimento.

(*Chiripa vuelve a lanzar la moneda cien veces, y ahora se obtienen 66 "caras".*)

Profesor: Han ocurrido 66 "caras". Tiene la palabra la defensa.

Stadi Shka: Sin comentarios, su señoría. ♣*Se le aguaron los ojos.*♣

Profesor: ¿La señorita fiscal desea decir algo más?

Ana Liza: No, señoría. ♣*Sonriente como nunca.*♣

Profesor: No habiendo lugar a más discusión, se levanta la sesión temporalmente y mientras tanto, el jurado entra a deliberar para emitir el fallo.

(Después de cinco minutos de deliberar, aparece el señor juez con la decisión del jurado.)

Profesor: Por favor, póngase de pie el acusado. (*Tahuro se levanta*.) El jurado ha decidido declarar como culpable al acusado. Se le condena a pagar una multa de 32 almuerzos y entra en prueba disciplinaria. ♣*No puedo creerlo; deben estar cometiendo un error.*♣

a. ¿Son las pruebas contra Tahuro contundentes? Explique su respuesta.

b. ¿Qué tipo de errores podría cometer el jurado al tomar una decisión sobre la acusación que recae sobre Tahuro? Considere ambos casos: cuando es declarado culpable, y, cuando se declara que no se encontró evidencia de que sea culpable.

c. Si usted obtuviera 60 "caras" en un experimento como el que se realizó en el juicio contra Tahuro, ¿qué diría: "la moneda está cargada" o "la moneda es legal"? Explique su respuesta en términos probabilísticos.

d. Explique por qué la probabilidad de obtener 40 ó menos "caras", ó, 60 ó más "caras", suponiendo que la moneda es legal, al lanzar una moneda cien veces es de 0,05.

e. La fiscal Ana Liza dice: "La probabilidad de obtener 60 ó más caras en 100 lanzamientos de la moneda, si ésta es legal, es de 0,025". ¿Es correcta esa afirmación? Explique.

f. Suponga que la moneda de Tahuro realmente está cargada y que se obtienen 55 "caras" en el experimento de Ana Liza. ¿Qué decisión tomaría usted como jurado? ¿Sería justa su decisión?

g. En la subsección titulada "Búsqueda del criterio y final del juicio" usted propuso un criterio para juzgar la legalidad de la moneda de Tahuro. Determine cuál habría sido el veredicto, aplicando la norma que usted dio. Explique su respuesta. Determine qué tan amplia es su norma, en términos de probabilidad.

h. Si la norma acogida hubiera sido: Si salen menos de 35 caras, o, si salen más de 65 caras, entonces la moneda de Tahuro está cargada, ¿cuál habría sido el veredicto? ¿Qué probabilidad hay en ese caso de declarar culpable a Tahuro, no siéndolo?

i. Sabiendo que se obtuvieron 62 caras al lanzar la moneda cien veces consecutivas, construya un intervalo de confianza del 95% que le permita decidir acerca de la culpabilidad de Tahuro.

j. Suponga que la sospecha hubiera sido: La moneda de Tahuro está cargada a favor de "cara". En ese caso, no tendría sentido que la norma contemplara la posibilidad de condenar a Tahuro si salieran menos de a caras. Suponga entonces que el criterio fuera: si salen más de b caras, donde $50 < b < 100$, entonces, la moneda es ilegal. Se quiere que este criterio excluya sólo un 5% del total de los resultados posibles, ¿cuál debe ser el valor de b? Explique su respuesta.

Formalización de los conceptos

Hasta el momento se ha hablado de manera intuitiva acerca de conceptos tales como hipótesis nula, hipótesis alternativa y criterio de decisión. En lo que sigue trataremos de formalizar tales conceptos. Adicionalmente, se mencionarán los errores que se pueden cometer al tomar una decisión.

Las hipótesis

El juicio anterior presentó varios de los elementos de una prueba de hipótesis estadística. Veamos:

- Una sospecha acerca de la cual se desea realizar una investigación, la cual en estadística, se conoce como la *hipótesis de investigación*.
- Un principio para juzgar al acusado, en el que se supone que éste es inocente hasta que se demuestre lo contrario. En estadística, este supuesto es lo que se conoce como *hipótesis nula*, donde la palabra "nula" viene de nulidad, que en el caso del juicio indica simplemente que la moneda no está cargada. Como siempre existe la posibilidad de tener que rechazar la hipótesis nula, ésta siempre se enfrenta con otra hipótesis, conocida como *hipótesis alternativa*.

Criterio y decisión acerca de la hipótesis nula

Así como en un juicio se debe tomar una decisión acerca del acusado: declararlo culpable o inocente, también en estadística inferencial debe tomarse una decisión acerca de una hipótesis nula: rechazarla o no rechazarla. En un juicio, si se declara culpable a un acusado es porque se encontraron pruebas suficientes para no creer en su inocencia; en estadística, si se rechaza una hipótesis nula es porque se encuentran resultados significativamente diferentes a lo que

debería ocurrir si la hipótesis nula fuera cierta. Por otro lado, si en juicio se declara inocente al acusado es porque no hubo pruebas suficientes para alegar su culpabilidad; en estadística, el no rechazo de una hipótesis nula quiere decir que los resultados no fueron significativamente diferentes de lo que se esperaba, bajo la suposición de que la hipótesis nula era cierta.

En todo caso, siempre se debe establecer previamente un criterio para decidir acerca de la hipótesis nula, por lo cual se requiere:

- Un proceso con el que se trata de determinar la inocencia o culpabilidad del acusado, comparando el comportamiento de interés con lo establecido por las leyes en los códigos. En estadística este proceso se conoce como la determinación de la *región de rechazo* de la hipótesis nula.

Consecuencias de una decisión

Existen cuatro situaciones posibles originadas por la decisión de un jurado con respecto a la situación real y verdadera del acusado en cuanto a su culpabilidad, a saber:

- Declarar culpable al acusado, siendo éste inocente. En este caso, el jurado comete un error.
- Declarar culpable al acusado, siendo éste culpable. En este caso, el jurado toma una decisión correcta.
- Declarar inocente al acusado, siendo éste inocente. En este caso, el jurado toma una decisión correcta.
- Declarar inocente al acusado, siendo éste culpable. En este caso, el jurado comete un error.

También en estadística, al rechazar o no una hipótesis nula hay cuatro situaciones posibles con respecto a la correcta o incorrecta toma de decisión.

a. Complete el cuadro que se presenta a continuación, indicando en cada caso si la decisión es correcta o si se comete error.

		Situación real	
		La hipótesis nula es cierta	La hipótesis nula no es cierta
Decisión	Rechazamos la hipótesis nula		
	No rechazamos la hipótesis nula		

b. Explique el paralelo que se hace entre la decisión de un jurado y la decisión acerca de una hipótesis nula.

Al tomar una decisión acerca de una hipótesis nula es posible cometer el error de rechazarla siendo cierta ó de aceptarla, siendo falsa. En el primer caso, el investigador puede controlar la probabilidad de cometer error, puesto que es él quien fija un límite a partir del cual rechaza la hipótesis nula. En el segundo caso la situación es más complicada pues el investigador no tiene bajo control este error; por eso, cuando las pruebas no presentan evidencia que permita rechazar la hipótesis nula, la conclusión **no** es que se debe aceptarla; la conclusión es mucho más débil: es simplemente no rechazarla.

Proceso de las pruebas de hipótesis

Una vez que se tienen las ideas que apoyan la metodología de pruebas de hipótesis, vamos a presentar, de manera general, el procedimiento —paso por paso— que se sigue al realizar una prueba de hipótesis. Son siete los pasos que mencionaremos y después describiremos:

- Identificación del tipo de problema
- Planteamiento de las hipótesis
- Selección de la herramienta de análisis
- Selección de un modelo teórico
- Análisis del nivel de significación del resultado
- Toma de la decisión
- Presentación de conclusiones

Identificación del tipo de problema

En general, los problemas de inferencia estadística plantean el estudio de uno o más parámetros.[35] Entre los parámetros que se estudian con mayor frecuencia se pueden mencionar los siguientes:

- El valor de la media de una variable en la población. Parámetro que denotamos con la letra griega μ.
- El valor de la diferencia de medias de una misma población o de poblaciones diferentes. Denotamos esta diferencia de medias como: $\mu_1 - \mu_2$ ó $\mu_x - \mu_y$ ó con otros subíndices diferentes a los números 1 y 2 o a las letras x e y.
- El valor de la varianza de una variable de la población. La notación más usual es σ^2.
- El valor de la proporción de una variable en la población. La notación más utilizada para este parámetro es la letra "P" mayúscula.

Asociados a estos parámetros existen *estimadores de parámetros*. En el cuadro que se presenta a continuación se hace un resumen de parámetros y estimadores de parámetros con su respectiva notación.

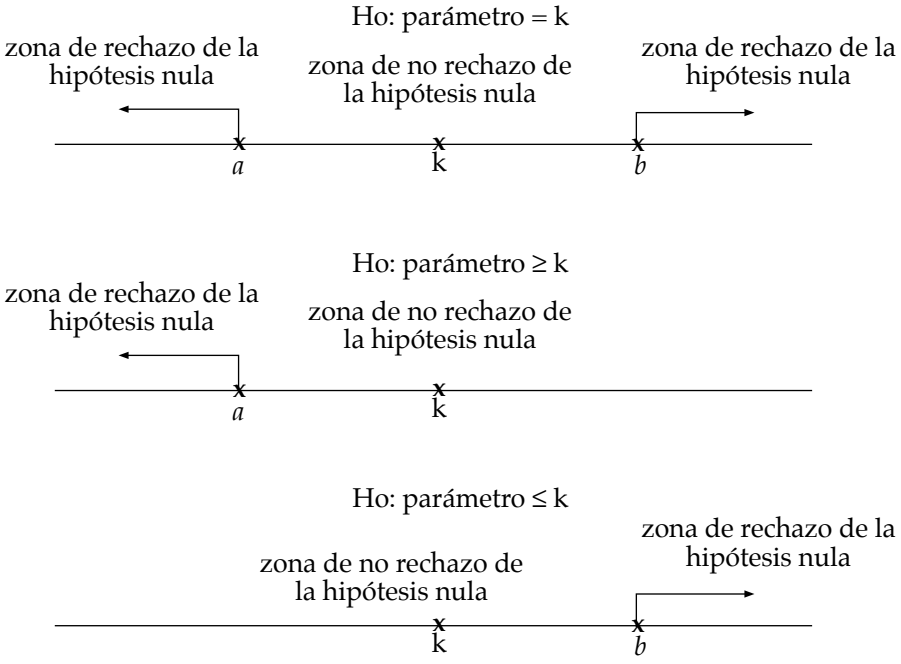
Parámetro	**Notación**	**Herramienta de análisis**	**Notación**
media poblacional	μ	media muestral	$\bar{x}$
diferencia de medias poblacionales	$\mu_1 - \mu_2$	diferencia de medias muestrales	$\bar{x}_1 - \bar{x}_2$
varianza poblacional	σ^2	varianza muestral	s^2
proporción poblacional	P	proporción muestral	$\hat{p}$

En este capítulo estudiaremos problemas relacionados con pruebas de hipótesis para la media μ y en el capítulo siguiente estudiaremos problemas relacionados con la diferencia de medias $\mu_1 - \mu_2$.

35 También hay casos de pruebas de hipótesis donde lo que se estudia es la distribución de la población en sí misma y, por ejemplo, hay pruebas para determinar si la distribución de una población es o no normal.

Planteamiento de las hipótesis

La hipótesis nula. La palabra "nula" transmite la idea de "ninguna diferencia". Como regla general, debemos comenzar con la afirmación de que no hay razón para creer que la sospecha que se tiene sea verdadera. La hipótesis nula se expresa de alguna de las siguientes formas:



donde k es un valor real conocido.

La hipótesis alternativa. Al plantear esta hipótesis, usualmente, debe recordarse el propósito de la investigación: buscar evidencia que permita rechazar la hipótesis nula. Por lo general, la hipótesis alternativa coincide con la sospecha que se tiene y es la negación de la hipótesis nula. Para los tres casos mencionados anteriormente, las correspondientes hipótesis alternativas son:

Ha: parámetro $\neq$ k

Ha: parámetro $<$ k

Ha: parámetro $>$ k

En el caso de que la alternativa no indique ninguna dirección específica, es decir, sea de la forma: parámetro ≠ k, se dice que la *prueba de hipótesis es de dos colas.*

Selección de la herramienta de análisis

Por lo general, en estadística inferencial encontramos más de una herramienta para enfrentar un mismo tipo de problema. En nuestro caso, emplearemos herramientas del campo de la estadística conocido como "estadística paramétrica". Algunas de las herramientas más usadas en este campo son: la media muestral $\bar{x}$ para inferir acerca de una media poblacional μ, la diferencia de medias muestrales $\bar{x} - \bar{y}$ para inferir acerca de una diferencia de medias poblacionales $\mu_x - \mu_y$ y la varianza muestral s^2 para inferir acerca de una varianza poblacional.

Selección de un modelo teórico

Para cada herramienta de análisis tal como $\bar{x}$, $\bar{x} - \bar{y}$, ó, s^2 existe una distribución muestral asociada. Ya se trataron anteriormente casos de distribuciones muestrales asociadas a $\bar{x}$ y a $\bar{x} - \bar{y}$. En todos los casos de distribuciones muestrales que se explicaron se observaron características tales como:

- forma acampanada de la distribución
- a mayor tamaño de la muestra, la distribución tiende más fuertemente a la forma acampanada

Las dos razones anteriores explican en parte la selección de la distribución normal como modelo que simula el comportamiento de las distribuciones muestrales que vamos a trabajar en este texto. A pesar de que la distribución normal es un "buen modelo" no siempre es el más adecuado. Existen otros modelos que, en ocasiones, pueden simular mejor el comportamiento de una distribución muestral de $\bar{x}$; sin embargo, esa discusión no se hará en este texto. Por ahora haremos uso extensivo del modelo normal, pero el lector debe ser consciente de que existen ciertas limitaciones que se han dejado de lado.

En resumen, usaremos el modelo de distribución normal tanto para distribuciones muestrales de $\bar{x}$ como para distribuciones muestrales de $\bar{x} - \bar{y}$.

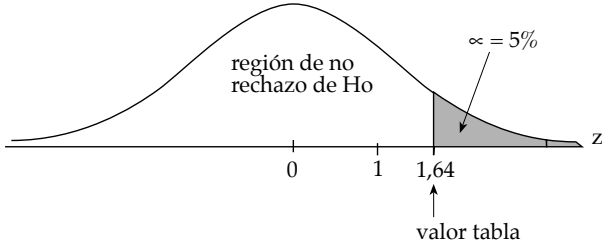
Análisis del nivel de significación del resultado

Una vez que se ha escogido un modelo para representar el comportamiento de la herramienta elegida para el análisis, determinar el nivel de significación es un problema de recetas de cálculo y de interpretación de la tabla de distribución del correspondiente modelo. El *nivel de significación* de una prueba de hipótesis, denotada por α, es la probabilidad que existe de rechazar la hipótesis nula. Hablando en términos de la gráfica, el nivel de significación es el área de la región de rechazo de la hipótesis nula. Generalmente, se utilizan los valores 5%, 2,5%, 1%, 0,5% para α.

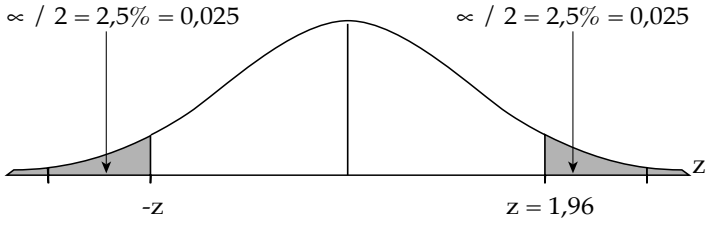
En el caso de que la prueba de hipótesis sea de dos colas, αes la suma de las áreas de las dos regiones de rechazo. Es decir, si $\alpha = x\%$ en una prueba de dos colas, entonces el área de cada una de las regiones de rechazo debe ser $(x/2)\%$.

Veamos un ejemplo del manejo del nivel de significación de una prueba de hipótesis:

En el modelo de la distribución muestral de medias se presenta el caso de una prueba de hipótesis de una cola, con un nivel de significación del 5%. En este caso, la hipótesis alternativa es: Ha: $\mu > k$. Buscando en la tabla, encontramos un valor aproximado para z de 1,645. Gráficamente sería así:



A continuación, en el modelo de la distribución de medias muestrales se presenta el caso de una prueba de hipótesis de dos colas con un nivel de significación del 5%. En este caso, la hipótesis alternativa es: Ha: $\mu \neq k$. El valor que se debería buscar en la tabla es el correspondiente a $\frac{\alpha}{2} = 2,5\%$. Buscando en la tabla encontramos z = 1,96. Gráficamente sería así:



Toma de la decisión

Una vez realizados los tres pasos que se explicaron anteriormente, la decisión acerca de la hipótesis nula, Ho, es simplemente cuestión de mirar en dónde cae el resultado de $\bar{x}$ con respecto al valor que buscamos en la tabla. La observación de la que se habla en la frase anterior se hace sobre el modelo de la distribución muestral correspondiente, para la cual su desviación estándar (error estándar) se definió en el capítulo titulado "Estadística inferencial". La gráfica donde ubicamos el valor de la tabla determina dos regiones: una, donde no rechazamos la hipótesis nula y la otra, donde la hipótesis nula debe rechazarse. Según dónde quede ubicado el valor del estimador muestral habrá que no rechazar o sí rechazar la hipótesis nula.

Presentación de conclusiones

En general, la toma de una decisión genera consecuencias y dudas que deben comentarse. Es importante no olvidar que la prueba de hipótesis no constituye una prueba contundente. Es posible cometer uno de dos errores. Rechazar la hipótesis nula siendo que es verdadera (error de tipo I, con probabilidad de cometerse igual a α) o no rechazar la hipótesis nula siendo que es falsa (error de tipo II; la probabilidad de cometer este tipo de error no es tan fácil de calcularse).

A continuación presentamos un ejemplo totalmente resuelto para ilustrar el desarrollo de una prueba de hipótesis.

Ejemplo: ¿Contaminación peligrosa en el centro de Bogotá?

Los habitantes de Bogotá viven preocupados por el aumento año a año de la concentración de bióxido de carbono (CO_2) en el centro de la ciudad. Se considera que un porcentaje de contaminación normal no supera el nivel del 24% de CO_2. En 1990 se contrataron especialistas que midieron el índice de contaminación. Los resultados que obtuvieron fueron los siguientes: en 36 días elegidos aleatoriamente durante el año se midió la concentración de CO_2 al medio día y se encontró una concentración media de 25% con una desviación estándar de 6%. De acuerdo con estos resultados ¿puede considerarse como peligroso el nivel de contaminación de CO_2 en el centro de Bogotá?

Identificación del tipo de problema. Se trata de un problema de inferencia acerca de una media poblacional. Podemos definir de manera formal este parámetro así:

> μ "media de los niveles de contaminación al medio día durante el año de 1990"

Planteamiento de las hipótesis. La hipótesis nula establece que "no pasa nada"; en este caso, el "no pasa nada" se debe interpretar como: la concentración media no alcanza un nivel peligroso, es decir, $\mu \leq 24$. Por tanto, la hipótesis nula es:

$$Ho: \mu \leq 24\%$$

La hipótesis alternativa debe plantear la posibilidad sobre la cual se tienen sospechas. En este caso, la sospecha es que el nivel medio de contaminación es peligroso, es decir, $\mu > 24\%>$. Por tanto, la hipótesis alternativa es:

$$Ha: \mu > 24\%$$

Como la hipótesis alternativa es de la forma ">" y no de la forma "≠", la prueba de hipótesis que se está planteando es unilateral o "de una cola" hacia el lado derecho.

Selección de una herramienta de análisis. La herramienta de análisis que se utilizó es la media muestral. Se encontró que $\bar{x} = 25\%$ con desviación estándar de 6%.

Selección de un modelo teórico. Ya hemos visto que el comportamiento de $\bar{x}$ se puede modelar de una manera aproximada con la distribución normal. Usaremos, entonces, ese modelo.

Análisis del nivel de significación. El resultado muestral estandarizado se establece con base en el cociente:

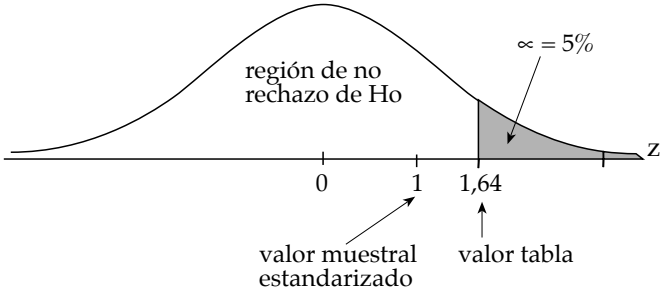
$$\frac{\text{media muestral - media poblacional de la hipótesis nula}}{\text{error estándar de la media muestral}}$$

En símbolos se obtiene:

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{25\% - 24\%}{\frac{6\%}{\sqrt{36}}} = \frac{1\%}{\frac{6\%}{6}} = 1$$

Valor de la tabla. Si elegimos como nivel de significación $\alpha = 5\%$, encontramos en la tabla un valor de 1,64.

Comparación de valores. La gráfica en donde se realiza la comparación se presenta a continuación:



Toma de la decisión. Se puede ver en la gráfica del punto anterior que el resultado muestral se ubica en la región de no rechazo de la hipótesis nula. Entonces, la decisión es no rechazar la hipótesis nula.

Presentación de conclusiones. Ya que no se rechazó la hipótesis nula se piensa que el nivel de contaminación de CO_2 no fue peligroso en 1990. Es posible que nos equivoquemos con esta decisión pues al no rechazar la hipótesis nula podemos estar "aceptando" una hipótesis que es falsa. Por eso, es mejor decir "no rechazamos" en cambio de decir "aceptamos".

Ejercicios

1.- Una planta química afirma que la producción diaria promedio de un cierto producto allí elaborado es μ =800 toneladas. Se tomó una muestra de la producción diaria de tal producto en 50 días y se obtuvo una media de 810 toneladas y una desviación estándar igual a 21 toneladas. ¿Presentan los datos suficiente evidencia para refutar la afirmación de la planta? (Use un nivel de significación igual a 0,05.)

2.- Muchos años de experiencia han demostrado que un examen de admisión de matemáticas produce una media de 55 con una desviación estándar de 9,3. Este año, los 100 estudiantes que presentaron el examen obtuvieron un promedio de 58. ¿Se puede afirmar que estos estudiantes están significativamente por encima del promedio?

3.- Para lanzar una campaña publicitaria sobre el ahorro de gasolina de un carro, se realiza un estudio para determinar el consumo de gasolina de carros del mismo tipo. Se tomó una muestra de 450 automóviles "corrientes" y se encontró que el consumo medio era de 26 km por galón y la desviación estándar de 4 km por galón. ¿Se podría afirmar que el consumo medio de gasolina es mayor que 24 km por galón?

4.- De un grupo de 2000 bebés, con edades entre dos y cuatro años, se tomó una muestra aleatoria de 100. Se encontró que el peso medio de la muestra es 8 kilos y la desviación estándar es 0,8 kilos. Con base en los resultados anteriores, ¿se puede rechazar la hipótesis Ho: μ =7,8

5.- A pesar de la riqueza lexical del español, la mayoría de los hispanoparlantes tienden a usar un número muy reducido de palabras diferentes con respecto al vocabulario existente.

Con el fin de implementar los cursos de redacción en español ofrecidos por una universidad bogotana, el departamento de idiomas desarrolló una investigación que medía el número de palabras diferentes sobre un total de 500 en 200 textos escritos por estudiantes de diferentes carreras de la universidad. Se encontró que la variable se distribuía casi normalmente con una media de 117 palabras y una desviación estándar de 14,3.

a. ¿Qué variable se toma en cuenta?

b. Un estudio anterior afirmaba que el número promedio de palabras diferentes usadas por los hispanoparlantes es menor que el 15% del total de palabras empleadas. ¿Lo encontrado en la investigación de la universidad corrobora la afirmación?

c. El departamento de idiomas creía que un número promedio aceptable de palabras diferentes era 150, para que los cursos no tuvieran que ser modificados. ¿La información presentada sugiere que debe haber cambios en los cursos de redacción del departamento?

Prueba de hipótesis sobre diferencia de medias para muestras independientes

Introducción

En el capítulo anterior se hizo una presentación general de la metodología estadística conocida como prueba de hipótesis y además se habló del caso particular en el que el parámetro sobre el cual se realiza la prueba es la media poblacional. En este capítulo vamos a hacer referencia al caso en el que la prueba de hipótesis se hace sobre la diferencia de medias, con el propósito principal de comparar las poblaciones de las cuales provienen las muestras con base en las cuales se va a inferir.

El proceso para la prueba de hipótesis sobre la diferencia de medias, es igual al empleado cuando se quiere validar estadísticamente una hipótesis sobre la media de una población. Sólo hay diferencias en los detalles: el estimador muestral no es la media, sino la diferencia de medias, por tanto la distribución muestral que sirve como fundamento teórico es la de diferencias de medias con la correspondiente fórmula para el error estándar. Por tal razón, este capítulo se desarrollará muy rápidamente. En primer lugar se proponen tres enunciados; se hace la caracterización y la solución de dichos problemas. Luego se hace un resumen de los aspectos relevantes para las pruebas de hipótesis sobre la diferencia de medias. Finalmente se plantean una serie de ejercicios.

Motivación

El libro de texto. Un profesor sostiene que el libro de texto empleado en un curso de matemáticas es uno de los factores que influyen y determinan la metodología de clase y por tanto, el libro adoptado incide en el desempeño de los

estudiantes en el curso. Para verificar su hipótesis decide realizar un experimento: durante un semestre desarrolla el mismo curso para dos grupos de estudiantes de la misma carrera en la misma universidad, empleando dos libros, el X y el Y, de características bien diferenciadas, uno en cada curso. Al final aplica el mismo examen y obtiene los resultados que se muestran a continuación:

	Grupo 1 (libro X)	Grupo 2 (libro Y)
tamaño		$n_y = 36$
media	$\bar{x} = 3,8$	$\bar{y} = 3,5$
desv. est.	$s_x = 0,46$	$s_y = 0,51$

¿Los resultados encontrados por el profesor apoyan su hipótesis de investigación?

¿Cuál vía es la más rápida? Stadi Shka tiene clase todos los días en la universidad a las 7 a.m., y para llegar generalmente toma la ruta A. En días pasados, Chiripa le sugirió que tomara la ruta B puesto que es más rápida para llegar a la universidad a esa hora. Stadi Shka quiso hacer un experimento antes de seguir el consejo de Chiripa. Durante dos semanas consecutivas tomó la ruta B, anotando cada día el tiempo empleado desde su casa hasta la universidad y las siguientes dos semanas registró el tiempo empleado en ir de su casa a la universidad por la ruta A. Los resultados fueron los siguientes:

	Ruta A	Ruta B
media	$\overline{x_A} = 37\text{min.}$	$x_B = 38,3\text{min.}$
desv. est.	$s_A = 3,346\,\text{min.}$	$s_B = 3,346\,\text{min.}$

¿La información que consiguió Stadi Shka con su experimento apoya el consejo que dio Chiripa?

Tratamiento para adelgazar. En un aviso publicitario dirigido a mujeres entre 25 y 35 años, que sufran problemas de obesidad, un instituto para adelgazar ofrece un cierto tratamiento que seguido, sin interrupción, durante tres meses, hace posible reducir 16 kilos o más. Algún interesado en el asunto sospecha que el anuncio exagera la bondad del tratamiento y se propone hacer averiguaciones al respecto. Para ello, consigue tener acceso al archivo de historias clínicas del instituto y obtiene al azar una muestra de 30 mujeres que han seguido el tratamiento y la información es la siguiente:

Paciente	Peso inicial (k)	Peso final (k)	Paciente	Peso inicial (k)	Peso final (k)
1	68	57	16	78	63
2	82	63	17	67	53
3	80	63	18	73	60
4	70	58	19	68	53
5	68	52	20	75	62
6	76	58	21	72	58
7	76	60	22	69	52
8	65	53	23	78	65
9	80	60	24	80	60
10	72	59	25	79	60
11	75	60	26	84	70
12	75	64	27	77	61
13	65	50	28	80	63
14	78	61	29	69	55
15	72	58	30	74	60

¿Presentan los datos, evidencia de que el anuncio sea exagerado?

Para cada uno de los problemas presentados, responda las siguientes preguntas.

a. Determine las variables que se están considerando. Además, diga de qué tipo son.

b. ¿Cree usted que se puede suponer que la variable de interés, la que se está midiendo, se distribuye normalmente en la población?

c. ¿Cuántas muestras se están considerando? ¿Son grandes o pequeñas? ¿Son independientes o relacionadas? Explique sus respuestas.

d. ¿Cuál es la pregunta que se pretende responder?

Caracterización y solución de los problemas

A continuación vamos a solucionar los tres problemas planteados. En cuanto sea posible seguiremos las etapas de solución que se propusieron en el capítulo anterior, a saber:

- Identificación del tipo de problema
- Planteamiento de las hipótesis
- Selección de la herramienta de análisis
- Selección de un modelo teórico
- Análisis del nivel de significación del resultado
- Toma de la decisión

El libro de texto

Identificación del tipo de problema. En el problema *El libro de texto*, la variable de interés es el desempeño de los estudiantes en el curso de estadística, medida a través de la calificación obtenida en un examen final; esa variable es cuantitativa continua. Otra variable considerada explícitamente en el enunciado del problema es el libro empleado en el curso. La función de esta variable es separar en dos grupos bien diferenciados a los estudiantes, situación que da lugar a dos muestras independientes entre sí: la medición que se hace sobre cualquiera de los estudiantes no depende de la medición hecha a otro estudiante, ni tampoco influye en la medición hecha a otro estudiante. Se puede considerar que las dos muestras son grandes. El problema pregunta si existe diferencia en los resultados obtenidos por los estudiantes, de acuerdo con el libro empleado como texto durante el curso. Es decir, la pregunta formulada sugiere comparar la calificación media de los estudiantes de los dos cursos en el examen final, con la intención de determinar si el libro empleado genera dos poblaciones de estudiantes que se comportan de maneras diferentes en lo que interesa para el caso. Por tanto, la solución del problema se hará realizando una prueba de hipótesis sobre el parámetro, diferencia de medias.

Planteamiento de las hipótesis. La sospecha que tiene el profesor es que existe diferencia en el desempeño de los estudiantes según el libro de texto que se siga en el curso. Con la intención de validar esa hipótesis estadística, él parte del supuesto de que no existe razón para que su sospecha sea verdadera (hipótesis nula), y se propone buscar evidencia que lo lleve a rechazar ese supuesto, de manera que tenga entonces que acoger como verdadera la negación del supuesto hecho inicialmente (hipótesis alternativa). Por tanto,

$$\text{Ho: } \mu_x - \mu_y = 0$$

Es decir, las poblaciones de las cuales se extrajeron las dos muestras, no se diferencian en cuanto a su media.

Como la sospecha del profesor sólo indica que hay diferencia en el desempeño de los estudiantes, pero no se refiere a cuál de los dos grupos puede tener mejor desempeño, entonces la prueba de hipótesis será de dos colas y la hipótesis alternativa será:

$$\text{Ha: } \mu_x - \mu_y \neq 0$$

donde μ_x representa la calificación media de la población de alumnos que utilizan el libro X, y μ_y representa la calificación media de la población de alumnos que utilizan el libro Y.

Selección de la herramienta de análisis. Puesto que el parámetro sobre el cual se va a inferir es la diferencia de medias $\mu_x - \mu_y$ entonces la herramienta que utilizaremos será la diferencia de medias muestrales $\bar{x} - \bar{y}$. Para este caso, el valor del estimador muestral es $\bar{x} - \bar{y} = 3{,}8 - 3{,}5 = 0{,}3$

Selección de un modelo teórico. El valor del estimador muestral, (0,3), pertenece a una distribución muestral de diferencias de medias cuyo comportamiento fue descrito en el capítulo titulado "Inferencia estadística". Para inferir en este problema vamos a utilizar lo que se sabe de esa distribución:

- si las muestras son suficientemente grandes, la distribución de diferencias de medias es aproximadamente normal, con media igual a cero y desviación estándar igual a

$$EE(\bar{x} - \bar{y}) = \sqrt{(EE(\bar{x}))^2 + (EE(\bar{y}))^2}$$

Como las muestras tienen igual tamaño y pueden considerarse grandes (n = 36), entonces el modelo que sigue esta distribución es el normal. Además, el error estándar de la distribución está dado por:

$$EE(\bar{x} - \bar{y}) = \sqrt{(EE(\bar{x}))^2 + (EE(\bar{y}))^2} = 0{,}1145$$

porque:

$$EE(\bar{x}) = \frac{\sigma_x}{\sqrt{n_x}} \approx \frac{s_x}{\sqrt{n_x}} = \frac{0{,}46}{6} \text{ y } EE(\bar{y}) = \frac{\sigma_y}{\sqrt{n_y}} \approx \frac{s_y}{\sqrt{n_y}} = \frac{0{,}51}{6}$$

Análisis del nivel de significación del resultado. Como ya se dijo anteriormente se realizará una prueba de hipótesis de dos colas. El nivel de significación que se utilizará será $\alpha = 5\%$. Ese nivel de significación determina entonces en el modelo dos puntos críticos que son los que delimitan las regiones de rechazo y de no rechazo de la hipótesis nula. Tales puntos críticos son: $z_1 = -1{,}96$ y $z_2 = 1{,}96$. Y el criterio que se usará para tomar la decisión se puede expresar así:

> Si el valor del estimador muestral, debidamente estandarizado, es menor que -1,96 o mayor que 1,96, entonces deberá rechazarse la hipótesis nula.

Por tanto, procedemos a estandarizar el valor del estimador muestral, para obtener lo que se llama el estadístico de prueba:

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{EE(\bar{x} - \bar{y})} = \frac{0{,}3 - 0}{0{,}1145} = 2{,}62$$

La diferencia de medias muestrales —0,3—, corresponde al puntaje z = 2,62 en el modelo de la distribución muestral asociada al problema.

Toma de la decisión. Puesto que $z = 2{,}62 > 1{,}96$, entonces se debe rechazar la hipótesis nula, es decir, debe rechazarse el supuesto de que no hay diferencia en el desempeño de los estudiantes que siguen los libros X y Y. Por tanto, el profesor encontró evidencia que le ayuda a sustentar su hipótesis. Por supuesto, esto no es una demostración tajante. Existe la posibilidad de que al rechazar la hipótesis nula estemos cometiendo un error. Sería un error del tipo I y la probabilidad de cometerlo es del 5%.

¿Cuál vía es la más rápida?

Identificación del tipo de problema. En el problema *¿Cuál vía es la más rápida?*, la variable de interés es el tiempo empleado por Stadi Shka para ir de su casa a la universidad antes de las 7 a.m.; esa variable es cuantitativa continua. Otra variable considerada explícitamente en el enunciado del problema es la vía que utiliza para hacer su recorrido. Esta variable da lugar a dos muestras independientes entre sí. Las dos muestras son pequeñas. El problema pregunta si existe diferencia en el tiempo necesario para hacer el recorrido, de acuerdo con la vía utilizada. Es decir, la pregunta formulada sugiere comparar el tiempo medio que se requiere para hacer el recorrido cuando se utiliza la vía A, con el tiempo medio requerido al utilizar la vía B, con la intención de determinar si la vía empleada genera dos poblaciones de tiempos que se comportan de maneras diferentes en lo que interesa para el caso. Por tanto, la solución del problema se hará realizando una prueba de hipótesis sobre el parámetro, diferencia de medias.

Planteamiento de las hipótesis. La sospecha que tiene Stadi Shka es que existe diferencia en el tiempo empleado para ir de su casa a la universidad a una determinada hora del día, según que vaya por la vía A o por la vía B. Y, en este caso, la sospecha incluye una determinada dirección porque se cree que el tiempo empleado al ir por la ruta B es menor que el que se emplea al ir por la ruta A. Se tiene que:

$$\text{Ho: } \mu_A \geq \mu_B \rightarrow \mu_A - \mu_B \geq 0$$

$$\text{Ha: } \mu_B < \mu_A \rightarrow \mu_B - \mu_A < 0$$

donde μ_A representa el tiempo medio de la población de tiempos empleados al ir por la ruta A en las condiciones del problema, y μ_B representa el tiempo medio de la población de tiempos empleados al ir por la ruta B.

Selección de la herramienta de análisis. Puesto que el parámetro sobre el cual se va a inferir es la diferencia de medias $\mu_A - \mu_B$ entonces la herramienta que utilizaremos será la diferencia de medias muestrales $\overline{x_A} - \overline{x_B}$. Para este caso, el valor del estimador muestral es:

$$(37 - 38{,}3) = -1{,}3$$

Selección de un modelo teórico. El valor del estimador muestral (-1,3) pertenece a una distribución muestral de diferencias de medias cuyo comportamiento fue descrito en el capítulo titulado "Inferencia estadística". Para inferir en este problema vamos a utilizar lo que se sabe de esa distribución:

- si las muestras son suficientemente grandes, la distribución de diferencias de medias es aproximadamente normal, con media igual a cero y desviación estándar igual a -0,35.

Como las muestras son pequeñas (n = 10), entonces el modelo que sigue esta distribución no es el normal. Por tanto, aunque la esencia del procedimiento es la misma que conocemos no podemos terminar la prueba de hipótesis.

Tratamiento para adelgazar

Identificación del tipo de problema. En el problema *Tratamiento para adelgazar,* la variable de interés es el peso de las mujeres que siguen el tratamiento; esa variable es cuantitativa continua. En este caso hay dos momentos en los que se mide la variable de interés: antes de iniciar el tratamiento y al finalizarlo. Lo anterior da lugar a dos muestras, pero no es razonable considerarlas independientes entre sí, puesto que seguramente la medición de los pesos del sujeto *i* antes y después del tratamiento tienen alguna relación. Las dos muestras pueden considerarse suficientemente grandes. El problema pregunta si puede considerarse la diferencia entre el peso final y el peso inicial (de quienes conforman la población del problema) igual o superior a los 16 kilos. Este problema es diferente a los dos anteriores: aunque también implica la comparación de los pesos medios de dos grupos, se puede transformar en un problema de una sola muestra en la que la variable de interés sea la diferencia de los pesos final e inicial de cada una de las 30 personas cuyos datos se tienen. Así, la solución del problema se hará realizando una prueba de hipótesis sobre el parámetro, media poblacional.

Al realizar las diferencias de los valores de cada una de las parejas de pesos, se obtiene lo siguiente:

Paciente	Peso rebajado (k)	Paciente	Peso rebajado (k)
1	11	16	15
2	19	17	14
3	17	18	13
4	12	19	15
5	16	20	13
6	18	21	14
7	16	22	17
8	12	23	13
9	20	24	20
10	13	25	19
11	15	26	14
12	11	27	16
13	15	28	17
14	17	29	14
15	14	30	14
$\bar{x} = 15{,}13$ kilos, $s_x = 2{,}46$ kilos			

Planteamiento de las hipótesis. La sospecha que se tiene es que el parámetro de interés (el peso medio rebajado por quienes siguen el tratamiento) es inferior a 16 kilos. Por tanto, la prueba de hipótesis que se realizará es de una cola y las hipótesis son:

$$\text{Ho: } \mu \geq 16$$

$$\text{Ho: } \mu < 16$$

donde μ representa el peso medio rebajado, en la población.

Selección de la herramienta de análisis. Puesto que el parámetro sobre el cual se va a inferir es la media poblacional μ entonces la herramienta que utilizaremos será la media muestral $\bar{x}$. Para este caso, el valor del estimador muestral es $\bar{x} = 15{,}13$

Selección de un modelo teórico. Puesto que la muestra es grande (n = 30), el modelo que sigue la distribución muestral de medias a la cual pertenece el valor del estimador muestral, (15,13), es el normal y el error estándar de tal

distribución está dado por $EE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \frac{2{,}46}{6} = 0{,}4491$

Análisis del nivel de significación del resultado. Como ya se dijo anteriormente se realizará una prueba de hipótesis de una cola. El nivel de significación que se utilizará será α=1%. Ese nivel de significación determina entonces en el modelo un punto crítico que delimita las regiones de rechazo y de no rechazo de la hipótesis nula. Tal punto crítico es $z_1 = -2{,}33$. Y el criterio que se usará para tomar la decisión se puede expresar así:

> Si el valor del estimador muestral, debidamente estandarizado, es menor que -2,33 entonces deberá rechazarse la hipótesis nula.

Por tanto, procedemos a estandarizar el valor del estimador muestral, para obtener el estadístico de prueba:

$$\frac{\bar{x} - \mu}{EE(\bar{x})} = \frac{15{,}13 - 16}{0{,}4491} = -1{,}9372$$

El peso medio rebajado, dado por la muestra —15,13—, corresponde al puntaje z = -1,9372 en el modelo de la distribución muestral asociada al problema.

Toma de la decisión. Puesto que z = -1,9372 > -2,33, entonces no es posible rechazar la hipótesis nula, es decir, no se encontró evidencia de que se esté exagerando la bondad del tratamiento. Por supuesto es posible que efectivamente sí se esté exagerando sobre la bondad del tratamiento, en tal caso, al tomar la decisión de no rechazar la hipótesis nula estaríamos incurriendo en un error de tipo II.

Resumen

Las características de los problemas que resolveremos en este capítulo son:

- La variable de interés es cuantitativa continua y se puede suponer que su distribución en la población es aproximadamente normal.
- Se tiene información de dos muestras independientes. De cada una de ellas se conoce su tamaño, su media y su desviación estándar.
- Las muestras son grandes y sus tamaños son iguales o similares.
- Se supone que las varianzas de las poblaciones de las cuales provienen las muestras son iguales.
- La pregunta del problema sugiere hacer una prueba de hipótesis sobre el parámetro, diferencia de medias, puesto que se deben comparar las poblaciones de donde provienen las muestras; dicho de otra manera, se pretende averiguar si las muestras tomadas pertenecen o no a la misma población.

El investigador tiene la sospecha de que una variable no se comporta de la misma manera en dos grupos de una población, y por tanto, que para esa variable habría que considerar dos poblaciones en cambio de una. Para validar su sospecha extrae de cada uno de los grupos de la población, una muestra aleatoria y lo que se propone es ver qué tan significativa es la diferencia que las muestras presentan en su comportamiento.

Parte del supuesto de que las dos muestras provienen de la misma población (lo cual se expresa esperando que la diferencia de sus medias sea un número próximo a 0), determina un criterio con base en el cual juzgará si la diferencia de las medias de las dos muestras es o no significativa y procede a aplicar el criterio. En caso de que encuentre que la diferencia de las medias muestrales es significativa, entonces dirá que encontró evidencia de que las dos muestras provienen de poblaciones diferentes.

La lógica que sigue en su razonamiento es: en caso de que las dos muestras provengan de la misma población, la diferencia en el valor de sus respectivas medias ha de ser un número muy cercano de 0; en caso de que sea un número muy mayor que 0 ó muy menor que 0, entonces podría pensarse una de dos cosas:

- Se tomaron elementos de la población que constituyen los casos más extremos en ambas muestras, lo cual es posible pero poco probable si se hace muestreo aleatorio.
- Los elementos de las dos muestras no pertenecen a la misma población.

La solución de este tipo de problemas se apoya en la distribución muestral de diferencias de medias, la cual sigue el modelo normal para los casos en que las muestras son grandes.

Ejercicios

1.- Con el fin de diferenciar dos tipos de colonias de bacterias que presentan idénticas características, excepto en lo que concierne al tiempo que tardan en reproducirse, un investigador escoge dos muestras de cada tipo de bacterias y las observa durante tres meses. El mide el tiempo que tarda cada una de las 40 bacterias de cada muestra en empezar a reproducirse. Al final obtiene que en la muestra uno, el número promedio en horas de tal tiempo es 66, con una desviación estándar de 7 horas. En la segunda muestra se observó un promedio de 64,1 horas y una desviación estándar de 6,8 horas. ¿Puede concluirse que se trata de tipos de bacterias diferentes?

2.- Se quiere comprobar la eficiencia de un medicamento para aumentar la estatura de las personas. Se tomó una muestra de 50 sujetos, se hicieron las mediciones correspondientes y se encontró que la estatura media de ese grupo era 1,65 metros y su desviación estándar era de 0,03 metros. Al final del tratamiento se midió la estatura de 47 de los sujetos que lo siguieron y se encontró una estatura media de 1,73 metros y una desviación estándar de 0,05 metros.

Suponga que usted quiere determinar si hay evidencia suficiente para hablar de la eficiencia del medicamento y que para ello va a realizar una prueba de hipótesis. Responda las siguientes preguntas:

a. ¿Va a realizar una prueba de hipótesis sobre la media de la población, o sobre la diferencia de medias de dos poblaciones? Justifique su respuesta.

b. ¿Cuál es la hipótesis nula?

c. Realice una prueba de hipótesis con un nivel de significación del 5%.

d. Exprese la conclusión correspondiente.

3.- Un agricultor desea probar un nuevo insecticida que, según sus fabricantes reducirá las pérdidas causadas por ciertos insectos. Para probar la afirmación, el agricultor utiliza el nuevo insecticida en 200 árboles y el insecticida estándar en otros 200 árboles. Los resultados fueron los siguientes:

	Nuevo insecticida	**Insecticida estándar**
Producción media por árbol (kg)	240	227
Varianza	980	820

¿Proporcionan los datos suficiente evidencia para concluir que los dos insecticidas son diferentes en calidad? Explique.

4.- Para explicar un cierto tema en dos secciones de un curso de matemáticas, el profesor empleó dos métodos diferentes de enseñanza (en la sección 1, empleó el método A y en la sección 2, empleó el método B) y luego aplicó el mismo examen en las dos secciones. La información obtenida se presenta en la siguiente tabla:

	Sección 1	**Sección 2**
número de estudiantes	49	49
calificación promedio obtenida en el examen	3,64	3,86
desv. est. de la dist. de calif.	0,45	0,38

Además, se sabe que en ambos casos la distribución de las calificaciones es aproximadamente normal.

a. Con respecto a la sección 1, ¿cuántos estudiantes perdieron el examen? (Es decir, ¿cuántos estudiantes obtuvieron en el examen 2.9 o menos?)

b. Con respecto a la sección 2, ¿qué calificación debe tener una persona en ese examen, si su calificación supera la calificación del 85% de la gente del curso?

c. Determine si los resultados presentados en la tabla muestran evidencia suficiente para afirmar que hay diferencia significativa entre los dos métodos de enseñanza, en cuanto a la efectividad para la enseñanza. (Haga una prueba de hipótesis al 5% de significación.)

d. Suponga que el semestre entrante se va a emplear el método 2 para enseñar el tema mencionado y se va a aplicar el mismo examen. Emplee los resultados obtenidos en la sección 2 (que fue donde se usó tal método) para estimar con una confianza del 90%, la calificación promedio de los nuevos estudiantes del curso en dicho examen.

5.- Un sicólogo afirma que la duración promedio del tiempo de cortejo es mayor antes de un segundo matrimonio que antes del primero. Basa su afirmación en el hecho de que el promedio para efectuar un primer matrimonio (en una muestra de 626 parejas) es de 265 días con una desviación estándar de 50 días; mientras que el tiempo promedio para efectuar un segundo matrimonio (en otra muestra de 626 parejas) es de 268,5 días con desviación estándar de 53 días.

a. Si usted va a realizar una prueba de hipótesis para probar la validez de la afirmación hecha por el sicólogo, ¿cuál es la hipótesis nula?

b. ¿Cuál es la hipótesis alterna?

c. ¿La prueba que usted va a realizar es de 1 cola? ¿Por qué?

d. Si el nivel de significancia es del 1%, ¿debe usted rechazar o aceptar la hipótesis nula? ¿Por qué?

e. ¿Es válida la suposición del sicólogo?

6.- En casos de pacientes epilépticos, en ocasiones, la agresividad alcanza niveles más altos de lo normal.[36] Un grupo de científicos está buscando un método para disminuir el grado de agresividad que alcanzan algunas personas que sufren epilepsia psicomotora. Estas personas pueden llegar, incluso, a cometer delitos cuando se hallan en estado epiléptico. Dichos científicos creen que los neurotransmisores noradrenalina y dopamina

36 La agresividad es un componente de la conducta normal que se libera para satisfacer necesidades vitales y para eliminar cualquier amenaza contra la integridad física y psicológica del organismo.

pueden ayudar a disminuir este fenómeno. Para encontrar una solución a su inquietud hicieron un experimento con un grupo de 30 personas epilépticas. A cada una de ellas se le sometió a tres sesiones de descargas eléctricas de 10 segundos cada una. En la primera sesión se registraba, por medio de un electroencefalograma, el grado de agresividad (medido en una escala entre 0 y 10). En la segunda sesión se les inyectaba noradrenalina después de la descarga y se registraba igualmente el grado de agresividad que presentaban. En la tercera sesión se les inyectaba dopamina y se procedía de la misma manera. Entre sesión y sesión había un reposo de 24 horas, para que cualquier efecto posterior que tuviera el neurotransmisor hubiera pasado.

A continuación se presenta el valor medio correspondiente a la agresividad de los 30 sujetos en cada una de las sesiones y la desviación estándar respectiva.

	media	**desviación estándar**
Sesión 1	8,35	0,709
Sesión 2	7,45	0,65
Sesión 3	8,8	0,64

a. Haga una prueba de hipótesis con un nivel de significación del 1% para determinar si existe diferencia significativa entre el grado de agresividad producida por descargas eléctricas y la producida por descargas eléctricas, con la intervención de la noradrenalina.

b. Haga una prueba de hipótesis con un nivel de significación del 1% para determinar si existe diferencia significativa entre el grado de agresividad producida por descargas eléctricas y la producida por descargas eléctricas, con la intervención de la dopamina.

c. ¿Qué puede concluir usted acerca de los resultados?

7.- La directora de un instituto de niños afásicos desea hacer un cambio en las terapias que reciben los niños. Quiere aplicarles un método nuevo que según revistas mexicanas ha traído muy buenos resultados. Sin embargo, no quiere iniciar ese nuevo proceso con todo el instituto sin estar segura, por lo menos, en un 90% de que traerá resultados positivos.

Decidió entonces hacer un experimento. Para ello, escogió al azar 60 niños y con ellos formó dos grupos de 30. El primer grupo siguió la rutina normal durante un mes y al segundo grupo, le aplicó el nuevo método de terapia durante el mismo tiempo. Al finalizar el mes, a cada niño se le hizo tres evaluaciones: una auditiva, otra de cognición y otra de lecto escritura. A continuación se presenta el promedio de los tres resultados obtenidos por cada niño:

Grupo A (control)									
30	33	28	30	32	33	28	29	30	32
33	28	30	33	29	34	32	29	33	31
30	32	31	30	29	30	29	32	31	28
Grupo B (experimental)									
46	32	24	21	31	32	29	36	37	48
23	49	21	33	38	48	29	33	48	49
24	34	38	46	41	47	38	34	46	25

¿Existe diferencia significativa entre los resultados obtenidos en los dos grupos? ¿Supera el método nuevo al método tradicional? Justifique su respuesta.

8.- Durante los últimos años, Colombia ha sufrido una de las más grandes escaladas terroristas de su historia. Este fenómeno tiene graves repercusiones no sólo en lo político sino también en el campo económico ya que debilita el turismo al mismo tiempo que se presenta una fuga de capitales y la inversión extranjera se ve en peligro.

En marzo de 1991 se realizó un estudio en dos de las ciudades más turísticas de Colombia y se observó que el número de turistas extranjeros jóvenes y los ingresos percibidos por las empresas hoteleras ha aumentado notablemente con relación al trimestre final del año pasado. Para dicho estudio se tomó una muestra 200 personas en ambas ciudades; se encontró una edad promedio de 24 años con una desviación estándar igual a 2,3 años. También se observó que, el ingreso diario promedio por persona, a las empresas hoteleras, era de 70 dólares con una desviación estándar igual a 4,2 dólares. En un estudio realizado en octubre del año pasado, se encontró que la edad promedio de los turistas era de 32 años con una desviación estándar de 2,5 años y el ingreso diario promedio por persona era de $60 con una desviación estándar de $4,12.

a. ¿Cuál es la población de estudio, cuál es la población de datos y cuál la muestra del estudio?

b. ¿Qué variables se están midiendo? ¿Cuál es el parámetro sobre el cual se va a inferir?

c. ¿Cuál es el objetivo de la investigación?

d. ¿Se podría afirmar que por el aumento del terrorismo, la afluencia de turistas jóvenes es mayor porque desean conocer de cerca la realidad colombiana? ¿Qué proceso debe llevar a cabo para validar la afirmación anterior? Explique detalladamente.

e. ¿Existe suficiente evidencia para afirmar que hoy en día se dispone de mayor cantidad de dinero para gastar por parte de los turistas extranjeros?

9.- El narcotráfico es uno de los temas que, sin lugar a dudas, ha originado gran controversia tanto en el ámbito nacional como en el internacional. A raíz de esto son muchos los intentos que han realizado varios presidentes y organizaciones a nivel estatal (DEA) para tratar de combatirlo, pero hasta el momento los resultados son poco satisfactorios. Existen muchas razones que llevan a pensar que sería mucho más "efectivo" legalizar la actividad del narcotráfico que seguirla combatiendo, pues lo único que se ha conseguido con los métodos usados hasta ahora es la pérdida de vidas humanas y el desmesurado crecimiento de organizaciones delictivas en todo el mundo. Entre los argumentos citados están: la reducción de criminalidad por parte de los adictos para "financiarse el vicio", la disminución de asesinatos de jueces y políticos que han optado por lanzar "políticas anti drogas" en el país, el ingreso de divisas que incrementarían el presupuesto nacional y permitirían, por ejemplo, fortalecer programas de salud y educación.

Basándose en los planteamientos anteriores, un estudiante de Ciencia Política quiso conocer la opinión de sus compañeros de carrera acerca de la legalización del narcotráfico, teniendo en cuenta la edad de ellos. Para el efecto realizó una encuesta a 27 de tales estudiantes y encontró que estaban a favor de la legalización del narcotráfico aquellos cuya edad promedio era de 21 años con una desviación estándar de 1,3 años. El politólogo buscó contrastar sus datos con los de un estudio similar realizado a 26 personas, un año antes, en el cual se encontró que el promedio de edad era de 25 años.

a. ¿Cuál es la población de estudio?

b. ¿Cuál es la variable que se está midiendo?

c. ¿Cuál es el objetivo del estudio?

d. Comente sobre la calidad del muestreo.

e. ¿Podrá afirmarse que la opinión de los jóvenes de la muestra con respecto al tema ha cambiado? Justifique su respuesta.

10.- Estudiantes de quinto semestre de Ciencia Política realizaron una investigación en el departamento de la Guajira para conocer qué tanto varían los gastos económicos en grupos indígenas dedicados a la explotación de sal, con respecto al nivel de ingresos mensuales por familia. Tomaron una muestra en 35 caseríos, obtuvieron un gasto promedio de $22.780 con una desviación estándar de $975. Otro grupo de politólogos quiso corroborar dicho comportamiento: en una muestra de 33 caseríos encontraron un gasto medio de $21.920 con una desviación estándar de $930.

a. ¿Qué variable se está midiendo? ¿De qué tipo es?

b. Mencione los pasos que daría usted como investigador social para determinar si la muestra que se tomó en el primer estudio viene de la misma población de donde provino la segunda muestra. Explique detalladamente.

c. Determine con un nivel de significación del 5% si las muestras obtenidas por los dos grupos de politólogos provienen o no de la misma población.

11.- Un factor que ha influido notablemente en la evolución de las lenguas que trajeron a América los conquistadores europeos es la introducción de rasgos lingüísticos africanos a tales lenguas gracias al desarrollo de la institución esclavista en el Nuevo Continente. En el caso concreto del Brasil, país con una mayoría de población descendiente de los antiguos esclavos africanos, se observa una diferenciación entre el Portugués Brasilero Popular (PBP), propio de las clases sociales bajas, y el Portugués Brasilero Estándar (PBE), propio de las clases sociales altas.[37]

Una investigación que pretendía analizar sociolingüísticamente las diferencias sintácticas entre los dos tipos de habla portuguesa midió el porcentaje de variaciones sintácticas estigmatizadas[38] sobre el total de estas estructuras contenidas en la conversación de una persona. Para esto se realizaron entrevistas a personas de clase baja y de clase alta. Los porcentajes, distribuidos de manera aproximadamente normal, tenían por media y desviación estándar las siguientes:

	PBP	**PBE**
media, $\bar{x}$	41%	27%
desviación estándar, s	11,8%	7%
tamaño de la muestra, n	78	35

a. Un lingüista que había hecho investigaciones en el Brasil encontró que no había diferencia significativa entre la sintaxis usada por los hablantes de PBP y los de PBE. ¿Es esta afirmación acertada?

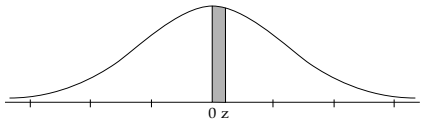
b. Otro estudioso de las lenguas criollas americanas afirmó que sí había diferencia, cualquiera que ésta fuera, a favor del PBP, es decir, que se comete mayor número de variaciones estigmatizadas en el PBP que en el PBE. ¿Es cierta esta afirmación?

c. Hace diez años, un lingüista encontró una diferencia igual a 3,7 entre los dos tipos de portugués. Los investigadores que hicieron el presente estudio esperaban que por la evolución que ha debido tener el PBP en este lapso, la diferencia entre los dos fuera la mitad de la encontrada hace diez años. ¿Se ha presentado la evolución que los investigadores suponían?

d. Los investigadores mencionados en el item anterior habían predicho el grado de evolución de la lengua. ¿Es válido hacer este tipo de predicciones? De acuerdo con su respuesta comente la aplicación de la estadística descriptiva e inferencial en estos problemas.

37 Guy, Gregory. "On the Nature and Origins of Popular Brazilean Portuguese". *Estudios sobre español de América y lingüística afroamericana*. Bogotá: Instituto Caro y Cuervo, 1989, pp. 227-230.

38 Una variación estigmatizada es una forma que difiere de la forma estándar y que por tanto es considerada incorrecta. Estas variaciones se asocian con el nivel socio-económico del hablante, es decir, que se asocia la forma "incorrecta" con una posición social baja.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	0.008	.0120	.0160	.0239	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3557	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4803	.4808	.4812	.4817	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Situaciones problemáticas

1.- Se lanza un dado 60 veces y se registra el resultado de la cara superior. Para cada lanzamiento, si el resultado es 1, usted paga $3; si el resultado es 2 ó 3, le pagan a usted $12; si el resultado es 4 ó 5 ó 6, usted paga $6.

a. ¿Estarían dispuestos a jugar el siguiente juego? [Supongan que una de las personas del grupo (el casino) propone el juego a la otra (el posible jugador)]. Justifiquen su respuesta.

b. Jueguen 60 veces y determinen cuál de las dos personas del grupo (el casino o el jugador) gana y cuánto. [Las cuentas deben ser claras].

La persona X aceptó jugar y los resultados que obtuvo se presentan a continuación:

1	6	4	4	5	3	1	4	1	5	5	1	2	3	1
6	4	6	4	3	5	5	6	3	4	3	1	6	4	3
2	3	3	1	6	3	4	4	2	6	2	1	5	3	1
6	5	4	2	4	6	6	4	6	1	5	1	3	1	6
2	1	4	1	5	6	1	1	6	1	2	2	6	1	1
2	5	6	2	4	6	6	2	3	1	4	2	3	1	4
5	2	6	1	3	6	5	1	5	3					

c. ¿Ganó o perdió? ¿Cuánto? [Las cuentas deben ser claras; esa claridad depende en parte de la forma como se organiza la información.] Sugieran una manera de procesar y organizar la información que se presentó.

d. Inventen alguna manera de analizar teóricamente si la definición de los pagos de este juego favorecen al casino o al jugador. (Para ello hagan los supuestos que crean convenientes).

e. Utilicen la misma manera para analizar la situación si se juegan 100, 500, 1.000, 50.000 100.000 veces. Para cada uno de los casos, determinen, en promedio, cuánto gana o pierde el jugador por cada juego que haga.

Se simuló el lanzamiento de un dado con un programa de computador. Los resultados obtenidos se presentan a continuación:

	Resultados					
# de lanzam.	1	2	3	4	5	6
100	14	13	18	19	20	16
500	74	73	83	100	82	88
1.000	158	169	160	172	173	168
10.000	1.610	1.734	1.667	1.644	1.642	1.703
50.000	8.356	8.451	8.379	8.401	8.230	8.183
100.000	16.730	16.661	16.775	16.748	16.636	16.450

f. Para cada uno de los seis casos que presenta la tabla, determinen la ganancia o pérdida del jugador. En promedio, ¿cuánto ganó o perdió el jugador en cada lanzamiento?

g. Comparen los resultados obtenidos experimental y teóricamente (preguntas **e.** y **f.**) y establezcan una conclusión.

2.- El juego que se va analizar ahora es el que se conoce como *Any Seven*. Se lanzan dos dados simultáneamente y el jugador apuesta a que la suma de los resultados en las caras superiores es 7. Si así ocurre, el jugador **recibe** (*recibe* no es sinónimo de *gana*) 5 veces la cantidad de dinero que apostó; de lo contrario, la pierde. Para poder jugar, el jugador debe "poner" –dar al casino– una cierta cantidad de dinero.

a. ¿Estarían ustedes dispuestos a jugar 60 veces Any Seven? ¿Y, 1.000.000 de veces? Justifiquen sus respuestas.

El jugador A jugó un cierto número de veces y los resultados que obtuvo se presentan en la siguiente tabla de doble entrada:

Resultados dado-2

Resultados dado-1	1	2	3	4	5	6	**Total**
1	4	2	2	5	3	5	21
2	4	2	2	1	7	6	22
3	3	3	4	4	1	2	17
4	7	1	3	4	2	6	23
5	2	2	3	2	4	4	17
6	5	2	2	2	3	6	20
Total	25	12	16	18	20	29	120

b. Supongan que el jugador A apostó cada vez $1. ¿Ganó o perdió? ¿Cuánto? Expliquen claramente cómo realizaron los cálculos para dar la respuesta.

c. Inventen una manera de establecer cómo se portó el azar con el jugador A. Expliquen claramente su respuesta.

d. Según la respuesta anterior, ¿cómo se portó el azar con el jugador A?

e. Si ustedes jugaran 1.000.000 de veces y cada vez apostaran $1, ¿esperarían ganar o perder? ¿Por qué?

f. Los jugadores B y C jugaron Any Seven, 50.000 y 60.000 veces respectivamente, apostando siempre $1. B perdió $9.000 y C perdió $9.560. ¿Con cuál de los dos jugadores se portó mejor el azar? Justifiquen la respuesta.

3.- Una editorial está interesada en promocionar cinco de sus últimos títulos y para ello piensa hacer, al azar, paquetes de tres libros diferentes y enviar-

los a determinadas librerías. El costo del envío de esos paquetes por correo depende del peso del paquete. De los cinco títulos, tres pesan igual (380 gramos) y dos pesan igual pero menos que los primeros (200 gramos). Para dentro de la ciudad, el correo tiene establecidos unos rangos de pesos con sus correspondientes precios; a continuación se presenta parte de esa información:

Rangos de pesos en gramos	Precios de envío ($)
...	...
[500, 700)	850
[700, 900)	950
[900, 1.100)	1.030
[1.100, 1.300)	1.100
...	...

a. ¿Cuál es el peso promedio de un paquete de los que piensa enviar la editorial a las librerías?

b. ¿Cuánto dinero espera gastar la editorial en el envío de 15 paquetes?

4.- Se lanza un dado y se observa el resultado de la cara superior. Si se obtiene 1, el jugador pierde $3; si se obtiene 2 ó 3, el jugador gana $12; si se obtiene 4, 5 ó 6, el jugador pierde $6. De esa manera, el juego está definido para favorecer al jugador.

a. Redefina las formas de pago para que el nuevo juego de azar no favorezca ni al casino ni al jugador. Justifique su respuesta.

b. Invente, con el lanzamiento de un dado, un juego de azar que sea atractivo para el jugador pero que su definición favorezca al casino.

5.- Represente gráficamente el comportamiento "ideal" (que se da en teoría) de los resultados posibles de:

a. Lanzar un dado un número muy grande de veces.

b. Lanzar dos dados simultáneamente, un número muy grande de veces.

6.- Explique de manera general cómo puede usted juzgar si en un determinado juego de azar, el azar se portó bien o mal con usted. (La explicación que se pide no debe basarse en ningún juego específico).

7.- Para analizar el comportamiento del azar en diversas situaciones se proponen dos enfoques: uno teórico y uno experimental. ¿Considera usted que todas las situaciones en las que está involucrado el azar se pueden analizar bajo ambos enfoques? Explique y dé ejemplos para justificar su respuesta.

8.- El Plan Nacional para el Desarrollo de la Microempresa –PNDM– desde su inicio ha venido atendiendo fundamentalmente empresas de los sectores económicos de manufactura, comercio y servicios. Las empresas objeto de atención del PNDM son aquellas que tienen hasta 10 trabajadores en el caso de la manufactura, y hasta 5 en los casos de comercio y servicios.

En el año 1993 el Banco Mundial podía desembolsar un préstamo de 50.000 millones de pesos para la financiación del PNDM, sin embargo, ese préstamo estaba condicionado a la justificación que la entidad diera con relación a las necesidades de capital de trabajo de las microempresas cobijadas. En consecuencia, la Dirección del PNDM se disponía a establecer esas necesidades y para ello diseñó una encuesta que debía aplicar a una muestra estratificada por sector económico y ubicación geográfica de las microempresas. La Dirección poseía información de segunda mano sobre la distribución del número de microempresas en todo el país, discriminada por las dos variables de estratificación. Esta se presenta en el siguiente cuadro.

Microempresas en Colombia en 1989

	Ubicación geográfica								
	Capitales			Resto del departamento			Total		
Sector	No.	%F	%C	No.	%F	%C	No.	%F	%C
Manufactura	40.705	62,63	11,21	24.293	37,37	9,37	64.998	100	10,44
Comercio	273.926	56,59	75,43	210.109	43,41	81,02	484.035	100	77,76
Servicios	48.504	66,04	13,36	24.939	33,96	9,62	73.443	100	11,80
Total	363.135	58,34	100	259.341	41,66	100	622.476	100	100

Fuente: DANE, censo económico multisectorial - 1990

Convenciones: %F: porcentaje fila; %C: porcentaje columna;
No: número de microempresas

%C = ((No. en la ubicación a buscar) / (No. Total columnas en esa posición)) x 100
%F = ((No. en la ubicación a buscar) / (No. Total filas en esa posición)) x 100

A continuación se presentan ejemplos de la forma como se calcula e interpreta la información de las columnas encabezadas como %F, y %C.

Ejemplo de %F:

- 62,63 se calcula haciendo el cociente de 40.705 sobre 64.998 multiplicado por 100 y significa el porcentaje de microempresas del sector manufacturero ubicadas en las capitales.

Ejemplo de %C:

- 11,21 se calcula haciendo el cociente de 40.705 sobre 363.135 multiplicado por 100 y significa el porcentaje de microempresas ubicadas en las capitales que pertenecen al sector manufacturero.

Ejemplo de total (%F):

- 58,34 significa el porcentaje de microempresas ubicadas en las capitales.

Ejemplo de total (%C):

- 77,76 significa el porcentaje de microempresas del sector comercio.

a. Con base en la información que da la tabla anterior y los ejemplos propuestos, construya las distribuciones de frecuencia (absolutas y relativas) de:

- la actividad económica de las microempresas,
- la ubicación geográfica de las microempresas.

b. A partir de las definiciones de problema y sistema social describa:

- el problema,
- el que considere el objetivo general,
- la población de estudio,
- las variables involucradas en la situación y las poblaciones de datos, que ellas generan.

c. Si por motivos de presupuesto, la Dirección del PNDM decidió que el tamaño máximo de muestra fuera 2.000. ¿Cómo conformaría tal muestra de manera que la población esté proporcionalmente representada en ella?

d. De acuerdo con la tabla, ¿cual sería la probabilidad de que al escoger una empresa del sector manufactura, ésta sea de alguna capital?

e. La encuesta aplicada a las 2.000 microempresas de la muestra con la que se trabajó dio la siguiente información: las microempresas del sector manufacturero en las capitales requieren, en promedio, 15 millones de pesos y las ubicadas en el resto del departamento, requieren en promedio 5 millones de pesos. Con base en la información anteriormente mencionada, ¿cuál es el valor del préstamo que esperaba hacer el PNDM a una microempresa de ese sector?

9.- El grupo económico del nuevo gobierno desea introducir cambios en las políticas fiscales del Estado con el fin de beneficiar a los microempresarios. Para ello, decidió hacer un análisis en la Administración de Impuestos de Bogotá dado que su recaudo representa el 60% del total nacional. Se escogieron 500 declaraciones de renta y patrimonio de personas naturales por estratos, teniendo en cuenta montos de renta declarados y sexo. Los datos obtenidos se presentan a continuación:

Distribución de personas que declaran renta

		Nivel de renta				
		NR1	NR2	NR3	NR4	NR5
Sexo	Hombres	87	85	66	43	22
	Mujeres	86	64	21	16	10

NRi: nivel i de renta por monto declarado

a. ¿Qué variables se consideran en el estudio y de qué tipo son?

b. Identifique las poblaciones y las muestras de estudio y de datos.

c. La muestra representa proporcionalmente a la población del estudio. ¿Cuál es la proporción de contribuyentes hombres de Bogotá, de nivel 2 ó 3?

d. Se ha establecido que el 10% del total de los impuestos pagados en Bogotá por personas naturales proviene del nivel 1 de renta. ¿Qué porcentaje de declarantes, según la muestra, hacen estos aportes?

e. Con base en la información de la muestra, represente gráficamente el número de declaraciones según nivel de renta.

El gobierno quiere eliminar el nivel de renta 1 (NR1) para favorecer (no gravar) a las personas que tienen menos ingresos. Por otro lado, se espera que para el próximo año se presente movilidad entre los diferentes niveles, es decir, se espera que en el nivel NR2 haya un aumento del 40% en el número de personas declarantes y un aumento del 8% en cada uno de los demás niveles (NR3, NR4, NR5).

f. Con base en la información de la tabla, represente gráficamente la nueva situación que se generaría con la eliminación del nivel 1 y con la movilización descrita anteriormente.

g. Si el gobierno decide no gravar los montos del nivel 1 de renta, para el año siguiente, ¿puede el gobierno esperar que haya más o menos declarantes teniendo en cuenta la movilidad descrita anteriormente?

h. De las siguientes afirmaciones, ¿cuáles son verdaderas para este caso? Explique.

- Es un experimento aleatorio.
- Se puede considerar una situación ideal.
- Aún con la incidencia del azar, se pueden estimar ciertos resultados.

10.- La compañía de seguros de vida *El Golpe* está estudiando la posibilidad de poner en el mercado un nuevo seguro de vida. Inicialmente y por razones de costos se venderá sólo en Bogotá. El gerente afirma que el seguro va a revolucionar el mercado de seguros de vida para personas que tienen 60 años. Se encuestó a 320 personas adultas de esta edad (38% hombres y 62% mujeres) y se encontró que el 62,55% de los hombres y el 57,7% de las mujeres estarían dispuestos a tomar este seguro si su valor no excede los $3.950.000 de cuota anual durante diez años y el pago en caso de fallecimiento es 11 veces la cuota anual al familiar más cercano. Según el censo de octubre de 1994 Bogotá tiene 7.123.446 habitantes, de los cuales 6,76% tiene 60 años a la fecha del estudio. Se conserva la proporción de sexos.

Para establecer la probabilidad de vida de las personas en cuestión se buscó en la resolución del 11 de abril del 94 de la Superintendencia Bancaria y se encontró la siguiente tabla de mortalidad de rentistas:

Tabla de mortalidad de rentistas. Experiencia I.S.S. 1980-1989

Sexo masculino		Sexo femenino	
edad	probabilidad de permanecer vivo	edad	probabilidad de permanecer vivo
51	0,99470	51	0,99550
52	0,99414	52	0,99507
53	0,99352	53	0,99462
54	0,99281	54	0,99409
55	0,99204	55	0,99351
56	0,99116	56	0,99287
57	0,99024	57	0,99221
58	0,98916	58	0,99151
59	0,98806	59	0,99073
60	0,98681	60	0,98992
61	0,98547	61	0,98892
62	0,98403	62	0,98772
63	0,98248	63	0,98628
64	0,98079	64	0,98464
65	0,97900	65	0,98274
66	0,97705	66	0,98075
67	0,97496	67	0,97868
68	0,97273	68	0,97648
69	0,97030	69	0,97032
70	0,96765	70	0,96922

a. ¿Cuál es el problema de estudio?

b. ¿Cuáles son la población y la muestra de estudio, y, la población y la muestra de datos? ¿Qué aspectos considera relevantes para la representatividad de la muestra?

c. La compañía quiere establecer cuánto dinero debe desembolsar el primer año por pago de seguros, atendiendo la variable sexo.

d. ¿Cuántos habitantes tenían 60 años en Bogotá a la fecha de estudio? ¿Cómo se discriminaban por sexo?

e. Según la tabla, ¿cuántas personas de 60 años mueren cada año?

f. Con base en los resultados de la muestra, ¿cuánto dinero debe desembolsar la compañía por pago de seguros el primer año?

g. Este mismo seguro generó utilidades el año anterior por $350.000.000. Este año, ¿sobrepasará esta cifra?

11.- Un grupo de politólogos, quería conocer la opinión de los bogotanos acerca del desempeño del entonces presidente de Colombia. Para ello se llevaron a cabo en Bogotá, en febrero de 1994, varias encuestas en las que se pedía calificar de 1 a 5 el desempeño del presidente, teniendo en cuenta el manejo de tres asuntos:

- manejo de la política exterior,
- manejo del problema del narcotráfico y
- manejo de la economía colombiana.

Para efectuar la encuesta se dividió la población por estratos teniendo en cuenta tanto el nivel socioeconómico como el sexo. En total se encuestó a 200 personas; las del estrato bajo fueron seleccionadas en el barrio "Meisen", las del estrato medio se seleccionaron en el barrio "Miranda" y las del estrato alto en el barrio "Santa Bárbara".

a. Identifique las poblaciones y muestras de estudio y de datos.

b. Para la división de la población por estratos y para la calificación del desempeño del Presidente se consideraron varias variables. ¿Cuáles fueron esas variables y de qué tipo son?

En la tabla siguiente aparecen tabulados los resultados de la encuesta para la calificación del manejo del problema del narcotráfico, discriminados según nivel socioeconómico y sexo:

	Estrato alto					Estrato medio					Estrato bajo				
Calificación	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Hombres	1	2	1	7	9	10	3	15	10	2	20	5	20	10	5
Mujeres	0	2	3	4	4	0	5	8	8	5	13	5	10	5	8

c. Complete la siguiente tabla (observe que las proporciones que se dan, se han obtenido teniendo en cuenta el número de hombres y de mujeres respectivamente).

Distribución de la calificación dada al manejo del problema del narcotráfico, según sexo

	Calificación				
	1	2	3	4	5
Hombres					0,1333
Mujeres		0,15		0,2125	

d. En la tabla anterior se discriminaron las proporciones según el sexo. Elabore una tabla similar a la anterior donde se discriminen las proporciones según el estrato socio-económico

e. En la muestra, ¿cuál es la proporción de personas de estrato alto que calificaron el manejo del narcotráfico con una nota de 4 ó 5? Compárela con la proporción de personas de estrato bajo, que también calificaron el mismo aspecto con una nota de 4 ó 5. ¿Qué se puede concluir?

f. Determine si las opiniones acerca del manejo del problema del narcotráfico, difieren menos por sexo que por estrato. Explique el criterio utilizado.

g. Si la encuesta se hiciera extensiva a una muestra de 35.000 bogotanos, ¿cuál cree usted que sería el número de personas que considerarían el manejo del narcotráfico con una importancia calificada de 3 o más? Justifique su respuesta.

h. Explique qué consideración fundamental se debe tener en cuenta para llegar a generalizaciones como la anterior.

Suponga que un compañero de clase le sugiere a usted, realizar 60 encuestas más del mismo tipo (guardando las mismas proporciones con respecto a estrato socio-económico y sexo) y le propone la siguiente apuesta: por cada persona que califique con 5 el manejo del narcotráfico, te pago $1.000; por cada persona que califique con 4 el manejo del narcotráfico, te pago $500; por cada persona que califique con 3 o menos el manejo del narcotráfico, tú me pagas $2.000.

i. ¿Estaría usted dispuesto a aceptar esta apuesta? ¿Cuánto dinero en total cree que ganaría o perdería? Justifique claramente su respuesta.

12.- Durante los primeros tres meses del presente año, casi a diario, los habitantes de la ciudad X han visto levantarse columnas de humo en diferentes zonas de la ciudad con el consiguiente olor a madera quemada de árboles de pino, vegetación nativa, matorrales y pastos que se chamuscan bajo el fuego.

La Oficina de Prevención de Emergencias de la ciudad, el Cuerpo de Bomberos, la Defensa Civil, el Ejército Nacional, y la Policía Metropolitana de la ciudad han visto reflejada esta situación en los datos de registro y reconocimiento de su Cuerpo de Bomberos, los cuales muestran la dimensión del problema ecológico a que se ha visto abocada la ciudad.

Puesto que el gobierno nacional debe prever el suministro anual de recursos para atender emergencias, quiere analizar la situación tal como se ha presentado en el primer trimestre del año, aceptando que eso representa una muestra de lo que puede ocurrir durante todo el año.

La siguiente tabla presenta los incendios ocurridos a la fecha.

Incendios forestales presentados en el primer trimestre del presente año

		Tipo de incendio	
		Menor	De gran magnitud
Ubicación geográfica	Cerros orientales	50	10
	Bosques de la parte alta del Silencio	40	5
	Cerros del Cable, Manjui, Conejera,	68	7

a. Defina (en sus palabras) el problema. Identifique el objetivo del estudio.

b. ¿Cuáles son las variables relevantes del estudio? Determine los valores que pueden asumir dichas variables y de qué tipo son ellas.

c. Identifique la población de estudio. Identifique la muestra de estudio, y las muestras de datos.

Con base en la información que da la tabla anterior, determine:

d. ¿Qué representa la cifra 68?

e. ¿Cuántos incendios han ocurrido en la ciudad X en el primer trimestre del presente año?

f. ¿Cuántos incendios han ocurrido durante el primer trimestre del presente año, en los bosques de la parte alta del Silencio y de gran magnitud?

g. ¿Cuál es la proporción de incendios ocurridos durante el primer trimestre del presente año en los bosques de la parte alta del Silencio, sabiendo que han sido de gran magnitud?

h. Haga una tabla de frecuencias relativas de:

- la ubicación geográfica de los incendios
- el tipo de incendios

i. Suponga que las proporciones del tipo de incendio se mantienen durante el resto del año. ¿Cuántos incendios de cada tipo se producirán durante el presente año? Si para combatir un incendio menor se necesitan dos helicópteros "baldebambi" y 15 en el caso de un incendio de gran magnitud, ¿cuántos helicópteros se esperaría utilizar para combatir una emergencia cualquiera?

13.- EL ICETEX está llevando a cabo una investigación para la adjudicación de becas y préstamos para estudios en el exterior. La investigación presenta dos tablas discriminadas según rangos de edades, del número de solicitudes que llegaron durante el año 1992, para becas y préstamos respectivamente así:

	Solicitudes de becas		**Solicitudes de préstamos**	
Rangos de edades	Frec. absoluta	Frec. relativa	Frec. absoluta	Frec. relativa
[20,22)	750	0,050	10.000	0,20
[22,24)	1.320	0,100	12.500	0,25
[24,26)	3.000	0,200	16.500	0,33
[26,28)	4.500	0,300	5.000	0,10
[28,30)	3.000	0,200	2.500	0,05
[30,32)	1.800	0,120	2.500	0,05
[32,34)	?	0,020	500	0,01
[34,36)	300	0,010	500	0,01
Total solicitudes	15.000		50.000	
Edad promedio	27,12 años		?	

El director del ICETEX analizó este informe y lo devolvió alegando que:

a. Falta información en la columna de frecuencia absoluta de solicitudes de becas.

b. Hay errores en la columna de frecuencia relativa de solicitudes de becas.

c. Falta el promedio de edades para solicitudes de préstamos junto con la explicación de cómo se calcula.

d. Falta la explicación de cómo se calcula el valor de la desviación estándar de edades para solicitudes de préstamos.

e. Quiere que las distribuciones se ilustren con gráficas.

f. Necesita saber cuál es la probabilidad de que en 1993, una persona de 30 o más años solicite préstamos, suponiendo que la distribución de préstamos para 1993 permanezca similar a la de 1992.

g. Quiere saber si para ambas distribuciones de edades (la de becas y la de préstamos) se puede asumir un modelo de distribución normal y por qué.

h. Espera que la distribución de edades para solicitudes de préstamos en 1993 se distribuya normalmente con la misma media y desviación estándar que en 1992. Bajo este supuesto necesita conocer:

- ¿Qué porcentaje de personas entre 27 y 35 años solicitará préstamos en 1993?
- ¿Cuál es la edad máxima del 20% de personas más jóvenes que solicitarán préstamos en 1993?

Usted debe responder las preguntas anteriores completando la información que haga falta corrigiendo los errores detectados y respondiendo a los interrogantes planteados por el director del ICETEX.

14.- El gobierno del presidente Samper tiene como una de sus prioridades el control de la inflación. Para ello, pretende lograr el compromiso de todos los colombianos –empresas, gremios, y en general, todo ciudadano que ofrezca servicios o productos– a través de su participación en lo que se ha

llamado Pacto Social. La implementación práctica del mismo puede resumirse en el hecho de que los incrementos en los precios no sea superior al 18%. Varias instituciones gubernamentales generaron diferentes medidas con el propósito de lograr la meta inflacionaria del gobierno. Algunas de ellas han sido consideradas "muy fuertes" o "exageradas" por los sectores directamente afectados. Por ejemplo, Fenalco consideró que se están dando los primeros pasos para la implementación de un control inadecuado de precios; el Consejo Gremial Nacional manifestó que para evitar que el comportamiento promedio de los precios exceda del 18% en 1995, no es necesario ni adecuado que la Comisión de Seguimiento monte esquemas de supervigilancia o que adopte sanciones a aquéllas empresas que en determinado producto presenten alzas de más del 18%. Ambos coinciden en que existen factores, como el incremento en los precios de los productos importados, el encarecimiento de insumos, que en determinados casos justifican que en algunos sectores el alza sea superior al 18%.

El Gobierno consciente de la necesidad de mantener la meta inflacionaria y los objetivos del Pacto Social, pero también consciente de que las críticas hechas por el sector económico a las medidas iniciales son razonables ha decidido replantear la presentación de la meta de inflación propuesta en dicho Pacto. En consecuencia, seleccionó una muestra de cuarenta productos y para ellos determinó el incremento que ha tenido su precio, con el fin de encontrar algún tipo de información que pueda utilizar en el nuevo planteamiento. Esta información se presenta a continuación:

Producto o servicio	Incremento (%)	Producto o servicio	Incremento (%)
Automóviles	18,3	Gasolina	19,3
Seguros	17,1	Papel periódico	21,0
Gaseosas	15,6	Pensión escolar	20,0
Jabón de tocador	17,2	Papel	23,0
Cerveza	15,8	Libros	20,6
Medias veladas	14,1	Leche	19,3
Café	18,9	Drogas	18,6
Transporte	19,0	Pescado	18,6
Azúcar	18,0	Shampoo	17,3
Pan	15,9	Arroz	15,9
Carne de res	20,0	Crema dental	17,3
Aceite	17,3	Frutas	18,2

Producto o servicio	Incremento (%)	Producto o servicio	Incremento (%)
Verduras	17,5	Mantequilla	19,5
Queso	18,9	Peluquería	17,5
Pasta	16,2	Discos	17,3
Servicio de luz	21,2	Cine	18,1
Servicio de teléfono	20,5	Arriendo o cuota	16,7
Servicio de acueducto	20,5	Servicios médicos	20,3
Artículos de aseo	19,7	Utiles escolares	20,9
Vestidos	18,7	Zapatos	16,5

a. ¿Cuál es el problema de estudio?

b. Determine la población y la muestra de estudio.

c. ¿Cuál es el atributo relevante de los elementos de la población por el que se interesa el problema? Es decir, ¿cuál es la variable relevante del problema? ¿De qué tipo es? ¿Qué población de datos genera esa variable?

d. Elabore una tabla de distribución de frecuencias de la variable que tenga seis intervalos; se quiere que el primer intervalo incluya el valor 14,1 y tenga una longitud de 1,5.

e. Presente gráficamente la información de la tabla anterior.

f. Utilice la información de la muestra para establecer alrededor de qué dato se podría afirmar que se han incrementado los precios.

g. Con base en la información de la muestra, un asesor del Gobierno manifestó que todos los datos están a dos o menos desviaciones estándar de la media. ¿Es acertada esta afirmación? Explique. ¿Sería razonable pensar que la distribución de esa muestra sigue el modelo normal? Justifique su respuesta.

h. Con base en las características que tiene una distribución normal, escriba un plan de acción que sirva para determinar si una distribución específica sigue ese modelo. Además, en caso de que no lo haya seguido para responder la pregunta anterior, sígalo.

Con base en la revisión hecha, el Gobierno estableció el tipo de aumentos que se pueden considerar como apropiados dentro de los lineamientos del Pacto Social, lo cual se puede resumir de la siguiente manera:

El incremento promedio de los precios será del 18,4%, aceptando una desviación estándar del 1.87%. Adicionalmente se supondrá que el incremento en los productos y servicios podrá variar de tal forma que siga una distribución normal.

i. Determine la proporción de bienes y servicios cuyo incremento oscila entre el 14,8% y el 17%, estando bajo el Pacto Social.

j. El Gobierno dará estímulos tributarios a aquéllas empresas cuyos productos o servicios presenten los incrementos más bajos. ¿Con qué incremento máximo se considerará entonces que el producto pueda generar deducciones, si se quiere dar estímulo al 20% de los productos que hayan tenido incrementos más bajos?

15.- Uno de los problemas de gran importancia en Colombia tiene que ver con el funcionamiento de la justicia. Por un lado, hay una gran cantidad de procesos pendientes de fallo: cada juez de la nación maneja más de 1.000 expedientes anualmente; en la actualidad, más de cuatro millones de procesos –que involucran a ocho millones de ciudadanos– están pendientes de fallo; además, un proceso penal demora en ser fallado de uno a diez años y, uno civil, puede durar entre cinco y siete años. Por otro lado, el costo de realizar los juicios es muy alto: Colombia es uno de los países que aporta más dinero a su sistema judicial; en América es el segundo país de más aportes, después de Costa Rica.

Con el fin de prever el presupuesto requerido para fallar los procesos, el Ministerio de Justicia debe determinar cuánto le cuesta al Estado realizar un juicio. Debido a la cantidad de expedientes, decide restringir el estudio a los procesos por corrupción, los cuales consideran faltas contra la moralidad y faltas contra la eficiencia pública.

La Procuraduría General de la Nación seleccionó al azar un juicio por cada uno de los tipos de procesos por corrupción presentados en el período de enero a junio de 1995, y encontró para ellos, los siguientes costos (sin considerar los sueldos ni los gastos por mantenimiento del sistema judicial).

Costo de juicios por corrupción	
Faltas contra la moralidad y la eficiencia pública	**Costo por proceso ($)**
Abuso de autoridad	1.480.000
Inmoralidad administrativa	1.790.000
Irregularidades en la contratación administrativa	1.200.000
Irregularidades en la prestación de servicios públicos	1.590.000
Mal manejo y uso indebido de bienes públicos	1.410.000
Irregularidades en el manejo del presupuesto público	1.320.000
Enriquecimiento ilícito	1.550.000
Extralimitación en el ejercicio de funciones	1.550.000
Omisión o negligencia en el ejercicio de funciones	1.610.000
Intervención en política	1.470.000
Irrespeto a la moral	1.350.000
Violación al régimen de inhabilidades	1.450.000
Faltas contra la administración de justicia	1.530.000
Violación de los derechos humanos	1.650.000

a. Defina en sus palabras el problema y el objetivo del estudio.

b. Determine la población y la muestra de estudio.

c. ¿Cuál es la variable de interés, de qué tipo es y qué valores puede tomar?

d. Elabore una tabla de frecuencias con 5 intervalos, cada uno de longitud 120.000.

e. Presente gráficamente la información de la tabla anterior.

f. ¿Se puede afirmar que la distribución que da la tabla sigue aproximadamente el modelo normal? Haga todos los cálculos que requiera para justificar su respuesta.

Con base en la información de la muestra, el Ministerio de Justicia en coordinación con la consejería presidencial para la Administración Pública y de acuerdo con el Plan de Transparencia han determinado que el presupuesto para cada proceso por corrupción deberá ser en promedio de $1.500.000. Además, suponen que el costo sigue una distribución normal con una desviación estándar de $150.000.

g. Con base en la información anterior, ¿qué proporción de procesos tendrán un costo entre $1.155.000 y $1.320.000?

h. El 5% de los procesos se considera "muy costosos". Si el proceso 8.000 hace parte de ese grupo de procesos, ¿cuánto es lo mínimo que podría costar?

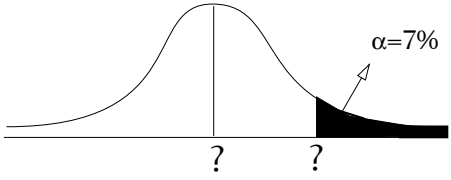
16.- Los gerentes de dos tiendas de comestibles encuentran que tienen exceso de confites importados. Los precios y las cantidades vendidas en ambas tiendas son idénticos. El gerente de la tienda A mantiene los confites en su lugar habitual, en tanto que el gerente de la tienda B los coloca durante un mes cerca de los mostradores donde los compradores pagan. Los gerentes registraron la cantidad de confites vendidos (en libras) diariamente durante un mes (30 días) y obtuvieron la información siguiente: en la tienda A se vendió en promedio 20,2 libras de confites con una desviación estándar de 1,8 libras. En la tienda B se vendió en promedio 21,9 libras con una desviación estándar de 2 libras. ¿El experimento que hicieron los gerentes de las dos tiendas da evidencia de una diferencia significativa en el número de libras de confites vendidas? ¿Puede concluirse algo con respecto al sitio de exhibición de los confites y las ventas que se logran? ¿Qué?

17.- A continuación se presentan algunos datos correspondientes a parte de la solución de un problema de inferencia estadística.

H_o: μ ? 6 $\quad \bar{x} = 6{,}45$

$s = 0{,}82$

H_a: μ ? 6 $\quad n = 49$



a. Redacte el enunciado de una situación que podría tener correspondencia con los datos que se dan.

b. ¿Qué interpretación y qué valores deberían tener los dos interrogantes presentados en la gráfica? Justifique sus respuestas.

c. Resuelva el problema redactado en la parte a., e interprete el resultado en términos del enunciado que usted propuso.

18.- A muchos estudiantes de este semestre, les ha gustado la moda del pantalón "bota campana". Felipe visitó esta semana 36 salones de clase de la Universidad X y encontró que varios estudiantes usaban este tipo de pantalón. Anotó algunos datos y con base en ellos hizo cálculos y encontró un promedio de 5 estudiantes por clase que usaban pantalón bota campana, con una desviación estándar de 2.

a. ¿Cuál es la variable de interés en este problema?

b. ¿Sobre qué parámetro se puede hacer inferencia en este problema?

c. ¿Qué tiene más sentido en este problema, realizar una prueba de hipótesis o un intervalo de confianza? Explique su respuesta.

d. Con base en los datos que da el enunciado y la respuesta que usted dio a la pregunta anterior, haga lo correspondiente (una prueba de hipótesis o un intervalo de confianza) e interprete el resultado obtenido en términos de la situación.

19.- Los directivos de cierta universidad afirman que, en razón de los mejores servicios que les ofrecen y de la supervisión más estricta de los trabajos, sus estudiantes completan el programa de doctorado en tiempo más breve que el usual. Basan esta afirmación en el hecho de que la edad media de los estudiantes al terminar el doctorado en el país es de 32,11 años mientras que los 36 estudiantes que culminaron el doctorado recientemente lo hicieron a una edad media de 29,61 años con una desviación estándar de 4,8 años.

a. ¿Cuál es la población de estudio? ¿Cuál es la muestra de estudio?

b. ¿Cuál es la variable relevante que considera el estudio?

c. ¿Considera usted que la variable que tomó el estudio es la mejor para trabajar la hipótesis de la universidad? Explique.

d. ¿Está usted de acuerdo con la afirmación de la universidad con respecto al tiempo que gastan los estudiantes haciendo el doctorado? Justifique estadísticamente.

e. Teniendo en cuenta la respuesta a la pregunta anterior, ¿considera usted que tiene sentido estimar la edad promedio que tienen los estudiantes de la universidad en cuestión al terminar el doctorado? Explique.

f. Estime la edad promedio que tienen los estudiantes de la universidad mencionada al terminar el doctorado.

20.- En los últimos dos meses del año en curso, los directivos de la fábrica de gaseosas "Quitased" han percibido en Bogotá una baja en sus ingresos y por ello sospechan que esa situación anuncia un cambio drástico en los ingresos de la empresa por ventas en la capital. Contratan un investigador para que analice la problemática y él decide hacer un estudio estadístico para establecer qué tan real es el problema detectado. Entre los aspectos que pretende estudiar figuran:

- preferencia actual por una determinada marca de gaseosa: Quitased, otra, ninguna
- razón principal para la preferencia: precio, sabor, facilidad para conseguirla
- consumo promedio semanal de la gaseosa preferida (en litros)
- cambio de preferencia por una determinada marca de gaseosa en los últimos cuatro meses: sí, no

De acuerdo con la Oficina de Planeación de Bogotá, los barrios están estratificados en tres grupos: alto, medio y bajo, información que tuvo en cuenta el investigador a la hora de recoger la información. El tomó una muestra aleatoria de 100 personas adultas y les aplicó una encuesta en la que preguntaba por los aspectos de interés. Encontró que 40 prefieren la marca Quitased, 45 prefieren gaseosas de otras marcas y 15 no prefieren ninguna marca. Para resumir la información de las 40 personas que prefieren la marca Quitased, elaboró las tablas 1 y 2 que se presentan a continuación:

Tabla 1. Distribución de frecuencias de las razones para preferir gaseosas Quitased según el estrato social

		Razón principal para la preferencia		
		Precio	**Sabor**	**Facilidad para conseguirlo**
Estrato social	Bajo	7	2	3
	Medio	5	8	7
	Alto	2	3	3

Tabla 2. Consumo promedio semanal (litros)

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

a. ¿Cuál es el problema de estudio? ¿Cuál es el objetivo del investigador? Defina la población y la muestra de estudio.

Las siguientes preguntas se refieren al grupo de encuestados que dijeron preferir la marca Quitased.

b. ¿Qué proporción de personas son de clase baja? Y, ¿de clase alta?

c. Si se elige al azar una persona, determine la probabilidad de que:

- ella prefiera la marca Quitased por razón diferente al sabor
- ella sea de estrato social bajo o medio y prefiera la marca Quitased por razón del precio
- ella prefiera la marca Quitased por razón del precio dado que es de estrato social bajo o medio

d. Represente gráficamente –de la manera más adecuada– los valores de la Tabla 2. Calcule la media y la desviación estándar del conjunto de datos. ¿Esta distribución se comporta de acuerdo al modelo normal? Explique.

e. Con base en la información encontrada en el punto anterior, estime la media del consumo promedio semanal con un nivel de confianza del 93%. Explique (**no lo haga**) de qué manera incide el hecho de variar el nivel de confianza del 93% al 97% sobre la precisión y la certidumbre de la estimación.

f. Hace un año, el mismo investigador realizó un estudio en el que involucro a 50 adultos residentes en Bogotá, de los tres estratos sociales y que consumían la marca Quitased, encontrando que la media del consumo promedio semanal era de 2,4 litros con una desviación estándar de 0,5 litros. El investigador sostiene que hace un año los consumidores de la marca en cuestión reportaban un consumo mayor. ¿Tiene razón? Explique. (Emplee un nivel de significación del 1%).

g. Del grupo de las 45 personas en la muestra que en la actualidad prefieren marcas diferentes a Quitased, sólo 5 dicen haber cambiado su preferencia en los últimos cuatro meses. El investigador sospecha que la correspondiente proporción en la población es mayor de 30%. ¿Qué respuesta da usted a la sospecha del investigador? Explique.

h. Utilice la información que usted obtuvo para hacer una análisis con respecto a los ingresos de la empresa.

21.- Un grupo de investigadores de comunicación social decide realizar un estudio para analizar diversos aspectos involucrados en el tiempo que los estudiantes universitarios de Bogotá en 1993, dedican diariamente a ver T.V. Además del tipo de programación que ellos prefieren, algunos de los aspectos que han considerado son los siguientes:

- Relacionar el tiempo que pasan viendo T.V. con el tipo de programación que ven.
- Comparar las creencias que tienen los programadores de T.V. con los resultados generados en una muestra de estudio.
- Determinar relaciones entre las preferencias por un tipo de programación y el canal en el cual se presenta.

a. Identifique el problema y la población de estudio.

b. Señale al menos dos objetivos que usted considere importantes en la investigación.

c. Clasifique los elementos considerados según el tipo de variable que le corresponda.

d. Sugiera una muestra de estudio que sea aleatoria, proporcional y representativa. Explique.

De un estudio realizado por otros investigadores en enero de 1989 se obtuvieron tres informes los cuales usted debe analizar cuidadosamente en las partes I, II y III.

Parte I

Para el Informe-I se tomó una muestra de 480 estudiantes universitarios de Bogotá con el fin de determinar la preferencia que ellos mostraban por la programación y los canales de T.V. Los resultados encontrados fueron los siguientes:

	Canal A	Canal B	Total
Noticieros	40	97	137
Telenovelas	92	51	143
Otros	105	95	200
Total	237	243	480

a. ¿Qué proporción de personas en la muestra ven noticieros?

b. ¿Qué porcentaje de personas en la muestra prefieren los noticieros y el canal B?

c. Si de la muestra se extrae aleatoriamente una persona, ¿cuál es la probabilidad de que ésta prefiera los noticieros dado que prefiere el canal B?

d. Si los eventos son: "ver noticieros" y "preferir el canal B". ¿Son independientes? Explique.

Parte II

Para el Informe-II se consideró el tiempo semanal que los estudiantes universitarios dedicaban a ver T.V. Se tomó una muestra de 100 estudiantes de la población y se les preguntó cuál es el tiempo (en horas) que dedican durante la semana a ver televisión.

a. ¿Qué tipo de representación gráfica emplearía para mostrar el comportamiento de la variable en la muestra? Explique su respuesta.

La siguiente tabla presenta la información de la muestra, agrupada en clases junto con su correspondiente frecuencia:

Distribución de frecuencias agrupadas del tiempo semanal dedicado a ver T.V.

Clases (horas)	Frecuencia
[0,0 - 3,5)	22
[3,5 - 7,0)	35
[7,0 - 10,5)	25
[10,5 - 14,0]	18

b. Con base en la tabla anterior calcule el tiempo medio semanal que las personas de la muestra dedicaban a ver T.V.

c. Determine en qué clase se ubica la mediana y la moda de la distribución de tiempos.

d. Considere la siguiente afirmación: "La variable tiempo semanal dedicado a ver T.V. en la muestra se comporta normalmente". Explique qué haría para determinar si esa afirmación es falsa o verdadera. (Sólo se le pide que explique. Usted no tiene que hacer cálculos).

Parte III

Con respecto al Informe-III, éste contenía dos tipos de información: resultados de algunos datos tomados de una muestra de estudio, y algunas creencias que tenían los directivos de la Universidad Nacional sobre la población de estudio. La siguiente tabla presenta la información:

Población de estudio: estudiantes de la Universidad Nacional	
Muestra de estudio: 300 estudiantes de la Universidad Nacional escogidos al azar	
Algunos datos de la muestra fueron:	**Creencias directivos de la U. Nacional:**
A. El 90% de los estudiantes ve T.V. todos los días.	**A.** El 80% de los estudiantes ve TV todos los días
B. Los 270 estudiantes que ven T.V. todos los días, la miran un promedio de 2.5 hora / día con una desviación estándar de 0.5 hora / día.	**B.** Las directivas no tenían ningún conocimiento previo acerca de este asunto.
C. 30% prefieren noticieros 40% prefieren novelas 30% prefieren otros tipos de programas.	**C.** 30% prefieren novelas 40% prefieren noticieros y 30% prefieren otros tipos de programas.
D. 90 estudiantes que tienen preferencia por los noticieros ven T.V. un promedio de 2,7 h / día, con una desviación estándar de 1,5 h / día. 120 estudiantes que tienen preferencia por las novelas ven T.V. un promedio de 2,4 h / día, con una desviación estándar de 1,2 h / día.	**D.** La gente que prefiere los noticieros ve menos televisión que la gente que prefiere novelas.

a. ¿Acerca de qué variables se está dando información en el cuadro anterior? ¿De qué tipo son?

b. Sobre dos de las variables anteriores, usted conoce métodos para hacer inferencia. ¿Sobre cuáles variables podría hacer inferencia?

c. ¿Qué tipo de inferencia estadística (prueba de hipótesis o intervalo de confianza) debería hacer si considera la parte **B** del cuadro anterior?

d. ¿Qué tipo de inferencia estadística (prueba de hipótesis o intervalo de confianza) debería hacer si considera la parte **D** del cuadro anterior?

e. ¿Acerca de qué parámetros haría usted la inferencia en los dos puntos anteriores (c) y (d)?

22.- En el año 1991 las exportaciones de banano colombiano llegaban a 25 millones de cajas; en el año siguiente decrecieron dramáticamente hasta llegar a 10 millones de cajas por razón de que las ventas a los Estados Unidos

bajaron sustancialmente. Con el fin de mejorar las ventas en el exterior se llevaron a cabo varias estrategias entre las cuales se pueden mencionar:

- La firma de un contrato de riesgo compartido con un cliente holandés muy importante –Velleman & Tass–, lo cual permitió gozar de los beneficios del mercado europeo.
- La modificación de la estructura financiera del sector, cambiando la deuda en pesos por deuda en dólares.
- La suscripción del acuerdo marco del banano con la Unión Europea que le otorga a Colombia, entre varios países, una cuota mínima de participación conocida como la cuota país.

Consciente de esta situación, la Unión de Bananeros (Uniban) decide hacer presencia en los mercados mundiales, específicamente en el mercado de la Unión Europea y en el mercado norteamericano el cual sigue siendo bastante atractivo y donde hay una fuerte competencia (Empresas como Chiquita –con un 26% de participación– y Dole –con un 30%–). Sin embargo, debido a restricciones de inversión se ve abocada a la decisión de definir en cuál de estos mercados debe enfocar más sus esfuerzos. Es así, como Uniban debe determinar cuál de los mercados es más atractivo en términos económicos, para lo cual deberá estimar el consumo del banano colombiano en las plazas mencionadas.

Para hacer la estimación de ventas se contrató un estudio con una compañía especializada. El estudio realizado seleccionó una muestra aleatoria de 200 clientes de la Unión Europea y encontró que el consumo promedio anual es de 32.500 cajas con una desviación estándar de 2.650 cajas. Así mismo, con base en una muestra de tamaño 150, de clientes del mercado norteamericano, se encontró que el consumo promedio anual de los clientes es de 33.205 cajas con una desviación estándar de 2.890 cajas anuales.

a. Defina el problema, la población y la muestra de estudio.

b. ¿Qué variables relevantes están involucradas? ¿De qué tipo son? ¿Qué poblaciones de datos determinan?

c. Estime el valor de la demanda promedio en número de cajas en la Unión Europea.

d. Un alto ejecutivo de la firma dice que el mercado norteamericano es más atractivo que el de la Unión Europea y que la probabilidad de equivocarse es del 1,0%. ¿Está usted de acuerdo? Justifique su respuesta.

23.- En las grandes ciudades del país, durante la última década, el número de padres de familia separados se ha incrementado sustancialmente, trayendo esto como consecuencia dificultades en el desarrollo emocional y mental de sus hijos. El Instituto Colombiano de Bienestar Familiar (ICBF) preocupado por el problema social que genera esa circunstancia contrató con la Universidad de los Andes el primer semestre del año 95 un estudio en Bogotá para detectar posibles causas de la separación de las parejas con hijos.

El grupo de investigadores de la Universidad aplicó una encuesta a 36 parejas que quieren seguir juntas y a 30 parejas que están pensando en separarse, todas con hijos. El contenido de la encuesta se presenta a continuación:

1) Tipo de unión
Religiosa _____ Civil _____ Libre _____

2) Quieren separarse
Sí _____No _____

3) La situación económica de su hogar les ha originado conflictos
Sí _____ No _____

4) Califique en una escala de 1 a 5 la tolerancia que tienen como pareja _____

5) Entre semana, ¿cuánto tiempo diario promedio comparten en pareja? (No incluya el tiempo de sueño) _____

En el grupo de parejas que piensan separarse se encontró que comparten en promedio 2,45 horas con una desviación estándar de 0,7 horas. Además, la calificación promedio que dieron a la tolerancia fue de 1,8 con una desviación estándar de 0,6.

Las parejas que quieren continuar juntas respondieron la pregunta 5) así:

2,3	3,7	3,9	3,7	2,6	2,0
2,0	3,0	3,2	2,8	1,2	1,9
3,0	4,2	3,3	4,1	1,8	2,2
3,3	3,0	2,8	3,6	3,4	2,5
2,5	2,0	1,98	1,7	1,3	1,4
2,3	2,5	2,6	2,3	2,4	2,7

Por otro lado, la calificación promedio que dieron a la tolerancia, las parejas que quieren continuar juntas, fue de 2,3 con una desviación estándar de 0,4.

Un estudio previo con parejas que quieren seguir juntas mostró que el nivel de ingreso que ellas tienen depende del tiempo diario promedio que comparten, de la siguiente manera:

Tiempo diario promedio compartido (h)	Nivel de ingreso ($)
2 o menos	4 millones
más de 2 pero menos de 4	3 millones
4 o más	2 millones

a. Defina en sus palabras el problema de estudio.

b. Establezca la población de estudio. Determine el objetivo del estudio.

c. Determine cuáles son las variables que se consideran en el estudio y cuáles de ellas estratifican la población. ¿De qué tipo son las variables?

d. ¿Cuántas muestras de estudio hay? ¿Cuántas muestras de datos hay? Menciónelas.

e. Describa la muestra de datos que se presenta en la tabla anterior. Determine si la variable se comporta de forma aproximadamente normal. Justifique su respuesta.

f. Utilice la información que relaciona el tiempo compartido con el nivel de ingreso de parejas que quieren permanecer juntas para determinar cuál es su nivel de ingreso esperado.

g. Existe la tendencia a pensar que la cantidad de tiempo compartido por la pareja influye en la calidad de su relación y por tanto, incide en el problema que se está abordando. ¿Cuáles son las dos variables del estudio involucradas en lo que afirma la tendencia? Exprese en sus palabras la tendencia mencionada y tradúzcala en términos de las correspondientes dos variables del problema.

h. ¿Se puede inferir del estudio de la Universidad de los Andes que el tiempo diario promedio de compartir es una causa de separación entre las parejas? Para responder, utilice la información recogida por los investigadores y siga lo propuesto en el punto anterior.

i. ¿Entre qué valores se estima que califiquen la tolerancia las parejas que quieren continuar juntas? ¿Existe una diferencia significativa entre la calificación promedio a la tolerancia que dan las parejas según que estén pensando en separarse o quieran seguir juntas? Para responder esta última pregunta puede utilizar la estimación que acabó de hacer.

j. Con base en el análisis hecho en las dos preguntas anteriores, concluya algo en relación con el objetivo del estudio.

24.- Con la fabricación de los primeros camperos la Willys de Colombia hace realidad un sueño que nació en Pereira en 1992. El Willys ha sido el jeep que durante 50 años ha recorrido todos los caminos colombianos, a pesar de que desde 1963 esta marca desapareció del mercado automotor. El nuevo Willys es una réplica del modelo 1954 con más capacidad de carga y motor más potente. Actualmente la compañía está produciendo 2 camperos mensuales pero sus proyecciones son mucho más ambiciosas. Sobre la comercialización de estos jeeps algunos voceros de las ensambladoras tradicionales consideran que áquellos entrarán a pelearse el mercado existente, mientras que otros, como Acolfa, señalan que los Willys ya tienen un mercado asegurado, especialmente en el eje cafetero. Pero si el mercado nacional resulta insuficiente las alternativas podrían estar en otros países del Tercer Mundo que tengan deficiencias viales similares a las de Colombia. Si bien la venta de carros ha caído en los primeros meses de 1996, los fabricantes esperan vender por lo menos un promedio de 40 por cada población. Para indagar cómo están sus expectativas, la empresa contrató una encuesta entre personas de 25 poblaciones del país y 24 poblaciones del Ecuador, sobre la aceptación (personas que contestaron que sí comprarían un campero atendiendo a sus características y a su precio.

Los datos de la tabla muestran los promedios de aceptación por población para cada marca. Fueron obtenidos de promediar los totales de personas, que comprarían esa marca, de las 25 poblaciones. Por ejemplo 39,2 es el promedio de personas que dijo que sí compraría un jeep Chevrolet.

	Colombia				Ecuador	
Marca	Promedio de aceptación por población	Desv. Estándar	Promedio de precio de venta (miles) por población	Desv. estándar	Promedio de aceptación	Desv. estándar
Chevrolet	39,2	2,7	1,900	920	36,6	2,1
Sofasa	24,3	1,6	17,200	530	18,7	1,4
CCA	38,1	1,9	18,500	878	26,2	1,0
Toyota	9,4	0,89	22,600	1140	11,8	0,99
Ford	5,9	0,64	18,100	826	17,9	1,1
Willys	35,5	2,9	18,300	770	33,1	2,0

Usted va completar el estudio iniciado con la encuesta.

a. Determine las variables relevantes del estudio y el tipo de cada una.

b. Con base en la información suministrada, establezca tres objetivos concretos de manera que, una vez desarrollado el estudio, permitan concluir algo con respecto al futuro de los camperos Willys.

c. Establezca un plan para el logro de los objetivos y llévelo a cabo utilizando los métodos de inferencia vistos en el curso.

d. ¿Cuáles son sus conclusiones sobre el mercado de los camperos Willys?

Referencias bibliográficas

Artigue, M. (1995). Ingeniería didáctica. En M. Artigue, R. Douady, L. Moreno, y P. Gómez (Eds.). *Ingeniería didáctica en educación matemática*. Bogotá: una empresa docente.

Batanero, C., Godino, J.D., Vallecillos, A., Green, D.R., Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *Int. J. Math. Educ. Sci. Technol*, Vol. 25, 4, pp. 527-547.

Burril, G. (1990). Quantitative Literacy: leadership training for masters teachers. En A. Hawkins (Ed.). *Training Teachers To Teach Statistics*, pp. 219-227. London: International Statistical Institute.

Chatfield, C. (1988). *Problem Solving. A Statistician´s Guide*. London, Chapman and Hall.

Ekeland, I. (1992). *Al Azar.* Barcelona: Editorial Gedisa.

Ellerton, N. & Clements, M. (1994). The Reshaping of Mathematics Education Research. En *ICMI Study 94 "What Is Research In Mathematics Education and What Are Its Results?"*. University of Maryland.

Fernández, F., Mesa, V., Gómez, P., Perry, P. (1993). *Estadística y Sociedad.* Bogotá: una empresa docente.

Garfield, J. (1995). How Students Learn Statistics. *International Statistical Review*, 63, 1, pp. 25-34.

Gómez, P., Mesa, V., Perry, P., Fernández, F., Gómez, C., Marulanda, I. (1993). *Matemáticas y Sociedad*. Bogotá: "una empresa docente" (Documento de trabajo)

Hawkins, A. (Ed.). (1990). *Training Teachers To Teach Statistics*. London: International Statistical Institute.

Kieren, T. & Pirie, S. (1994). Growth in Understanding: how can we characterize it and how can we represent it? *Educational Studies in Mathematics,* Vol. 26, pp. 171-181.

Kilpatrick, J. (1993). What constructivism might be in mathematics education. En J. C. Bergeron, N. Herscovics, y C. Kieran (Eds.). *Proceedings of the 11th International Conference for the Psychology of Mathematics Education*, Vol. 1, pp. 3-27. Montreal: Université de Montreal.

Lester, F. (1983). Trends and Issues in Mathematical Problem-Solving Research. *En Acquisition of Mathematics.* New York: Academic Press.

NCTM, (1981). *Teaching Statistics and Probability.* Reston, VA: NCTM.

NCTM, (1989). *Curriculum and Evaluation Standards for School Mathematics* Reston, VA: NCTM.

Perry, P., Fernández, F., Mesa, V., Gómez, P. (1990). *Matemáticas, Azar, Sociedad. Una introducción empírica a los conceptos de probabilidad*. Bogotá: "una empresa docente".

Phillips, J. (1992). *How to Think about Statistics*. New York: Freeman and Company.

Romberg, T. (1993). How one comes to Know: Models and theories of the learning of mathematics. En M. Niss (Ed.). *Investigations into assessment in mathematics education*. Dordrecht: Kluwer.

Rubin, A. & Rosebery, A. (1990). Teachers´ Misunderstanding in Statistical Reasoning; Evidence from a Field Test of Innovative Materials. En A. Hawkins, *Training Teachers To Teach Statistics*, pp. 72-101. London: International Statistical Institute.

Schoenfeld, A. (1992). Learning to Think Problem Solving, Metacognition, and Sense Making in Mathematics. En D. Grouws (Ed.). *Handbook of Research on Mathematics Teaching and Learning.* New York: McMillan, pp. 334-366.

Scholz, R. (1991). Psychological Research in Probabilistic Understanding. En R. Kapadia & M. Borovcnik (Ed.). *Chance Encounters: Probability in Education.* Amsterdam: Reidel, pp. 213-249.

Shaughnessy, J. M. (1992). Research in probability and statistics: reflections and directions. En D. Grouws (Ed.). *Handbook of Research on Mathematics Education.* New York: McMillan, pp. 465-494.

Shulte, A. & Smart, J. (Eds.). (1981). *Teaching Statistics and Probability. 1981 Yearbook*. Reston: NCTM.

Sobrino, M. (1994). Inferencia estadística en los bachilleratos. Revista *Suma, 17*, pp. 27-32.

Sorman, G (1991). *Los verdaderos pensadores de nuestro tiempo.* Bogotá: Seix Barral.

Steinbring, H. (1990). The Nature of Stochastical Knowledge and the traditional Mathematics Curriculum –some experience with in-service training and developing materials. En A. Hawkins (Ed.). *Training Teachers To Teach Statistics*. London: International Statistical Institute.

Tanur, J., Masteller, F., Kruskal, W., Lehmann, E., Link, R., Pieters, R., Rising, G. (Eds.). (1989). *Statistics: A Guide to the Unknown*. California: Wadsworth & Brooks.

Vere-Jones, D. (1995). The Coming of Age of Statistical Education. *International Statistical Review,* 63, 1, pp. 9-23.