4.9 Introducción a la Estadística: Lo fantástico de las estadísticas

Logros del aprendizaje

Después de estudiar esta sección podrás:

Resolver problemas de la medicina que envuelvan la efectividad de un medicamento

Explicar el significado de un intervalo de confidencia de 95%

Usar un intervalo de confidencia de 95% para interpretar resultados de encuestas

Explicar las dificultades en la construcción de una encuesta imparcial.

Por última vez volvemos de nuevo al objetivo de las cuatro secciones anteriores, lo cual fue formar herramientas para abordar los siguientes problemas:

> Las experiencias pasadas indican que cerca de la mitad de las personas que contraen cierta enfermedad mueren de dicha enfermedad. A base de esta experiencia, las personas en el campo de la medicina asumen que la probabilidad de muerte por esta enfermedad es de $\frac{1}{2}$. Se le administra una droga medicinal nueva a 24 personas que padecen la enfermedad y solamente 4 mueren de dicha enfermedad, mientras 20 se recuperan. ¿Es la droga nueva una droga efectiva para combatir la enfermedad o fue solamente el grupo particular (muestra) de 24 personas que respondieron de esta manera? ¿Cómo debe uno decidir si la droga nueva es efectiva o no?

Este problema fue mencionado inicialmente en la Sección 5 y ha sido repetido varias veces desde entonces, en varias secciones. Las herramientas están ahora a la mano. En el conjunto de personas que padecen la enfermedad, vamos a llamar a un sobreviviente un éxito y a uno que no sobrevivió un fracaso. Considera cada persona, en un grupo al azar de 24 personas que padecen la enfermedad, como un experimento en una secuencia de 24. Esta actividad lleva a la distribución binómica con $N = 24$ y $p = \frac{1}{2}$. Esta distribución binómica está estrechamente aproximada por una distribución normal con

$$\mu = Np = 24 \cdot \frac{1}{2} = 12$$

y una desviación estándar

$$\sigma = \sqrt{Np(1-p)} = \sqrt{24 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 2.449 \text{ (llevado a tres lugares decimales).}$$

Recuerda, en un grupo (muestra) de 24, la probabilidad que haya más de $\mu + 3\sigma$ éxitos es aproximadamente 0.001.

En este ejemplo (Figura 4.47),
$\mu + 3\sigma = 12 + 3(2.449) = 19.347$ (llevado a tres lugares decimales)

Figura 4.47

Esto es, sin la droga, la probabilidad que más de 19.347 personas sobrevivan en un grupo de 24 que padecen la enfermedad es aproximadamente 0.001. Sin embargo, usando la droga, encontramos 20 sobrevivientes en un grupo de 24 personas que padecen la enfermedad. Esto es más de 19.347. Si la droga no fuera efectiva, entonces, este resultado sería muy poco probable. Por consiguiente, a base de la información provista, la droga es declarada como efectiva.

1. **A base de las instrucciones dadas, ¿estamos seguros (100% certeza) que la droga es efectiva? Explica.**

2. **Las experiencias pasadas indican que cerca de dos terceras partes de las personas que contraen cierta enfermedad mueren de dicha enfermedad. Por consiguiente, las personas en el campo de la medicina asumen que la probabilidad de muerte por motivo de esta enfermedad es $\frac{2}{3}$. Una droga medicinal nueva se le suministra a 30 personas que padecen la enfermedad, y once mueren de dicha enfermedad. ¿Debe ser la droga nueva declarada como efectiva? ¿Estás seguro de tu respuesta? Explica.**

PHOTO

3. **Tienes una moneda que piensas es neutral, pero, no estás seguro. Esto es, piensas que si la moneda es lanzada, la probabilidad de obtener cara es $\frac{1}{2}$ y la probabilidad de obtener cruz es $\frac{1}{2}$. Si la moneda es lanzada muchas, muchas veces, entonces, las caras y cruces deben ocurrir con cerca de la misma frecuencia. Sin embargo, no estás seguro que esto ocurra. Esto es, no estás seguro que esta es una moneda neutral. Lanzas la moneda 40 veces y obtienes 10 caras. A base de este experimento, ¿dirías que la moneda es neutral? ¿Estás seguro de tu respuesta? Explica.**

   (*Pista*: Recuerda, con una curva de probabilidad normal, la probabilidad de una observación menor de $\mu - 3\sigma$ es aproximadamente 0.001)

Una curva de probabilidad normal tiene muchos otros usos que los mencionados en los problemas anteriores. Por ejemplo, los resultados de las encuestas, como reportados en la prensa o la televisión, usan frecuentemente esta curva excepcional. Las encuestas juegan un papel importante en nuestras vidas. Como un ejemplo, las encuestas pueden indicar la proporción de personas en un condado o distrito escolar que apoya el aumento en los impuestos escolares. Si las encuestas indican que la proporción es pequeña, entonces, los estudiantes, los maestros, los administradores y los partidarios van a aumentar sus esfuerzos para cambiar el parecer de las personas.

Enumera, por lo menos, dos otros usos para las encuestas con las que tú estás familiarizado.

El estudio de la *estadística* envuelve, entre otras cosas, las encuestas y las proporciones. Estas ideas ulteriores serán nuestro enfoque más importante en el resto de esta sección.

Antes de cada elección presidencial, los encuestadores tratan de seleccionar el ganador. Esto es, ellos tratan de adivinar qué proporción de la población votará por cada candidato. Claramente, encontrar la opción de todos los votantes sería un trabajo inútil. Como única alternativa, ellos encuestan, en el mejor de los casos, una muestra de unos cuantos cientos, o unos cuantos miles de personas con la esperanza que las proporciones en la muestra sean un buen estimado de las proporciones en la población total. Esto es un ejemplo típico de **inferencia estadística**. En particular, las características desconocidas de una población son inferidas de las características de una muestra observada en la población. Es interesante notar que el estudio de la estadística fue originalmente conocido como el estudio de la “aritmética política”.

Considera la siguiente situación:

Estás viendo el noticiero nocturno en la estación de televisión WMC (MC para **MATH** *Connections*). Se acerca una elección importante en el estado de Euphoria. Hay 200,000 votantes en el estado. La estación de televisión ha acabado de llevar a cabo una encuesta entre 1,000 votantes y encontró 600 que escogieron el candidato demócrata sobre el candidato republicano. Las aseveraciones siguientes son hechas por el anunciante: "Nuestra encuesta indica que el candidato demócrata tiene el apoyo del 60% de las personas en el estado. Este resultado de 60% tiene un margen de error de 3%".

PHOTO

1. **¿Cómo interpretas estas aseveraciones del anunciante? ¿Está de acuerdo tu interpretación con la de los otros estudiantes en la clase?**

2. **Si la estación de televisión condujera una encuesta entre 1000 votantes registrados, ¿piensas que exactamente el 60% de las personas en la muestra apoyarían al candidato demócrata? ¿Piensas que la proporción de personas que apoyan al candidato demócrata en la muestra nueva estaría cerca al 60%? Explica.**

3. **Supongamos que la estación de televisión ha conducido una encuesta entre 500 votantes y encontró 300 apoyando al candidato demócrata. ¿Piensas que el anunciante hubiera hecho las mismas aseveraciones. ¿Por qué?**

4. **Supongamos que la estación de televisión ha conducido una encuesta entre 10 votantes y encontró 6 escogiendo el candidato demócrata sobre el candidato republicano. ¿Piensas que el anunciante hubiera hecho las mismas aseveraciones? ¿Por qué?**

PHOTO

La proporción actual P de los votantes que apoyan el candidato demócrata es desconocida (hasta el día de las elecciones). Es casi imposible que $P = 0.60$, lo cuál es el resultado de la encuesta televisiva (recuerda que 60% = 0.60). Podemos ciertamente estimar 0.60 como un estimado de P y escribir

$$P = 0.60 \pm \text{algún error.}$$

Nuestro problema ahora es obtener una idea de "algún error".

Si tomamos muchas muestras al azar de 1,000 votantes y enumeramos las proporciones de personas en cada muestra que apoyan al candidato demócrata, la lista se parecería a la siguiente.

0.60 (nuestra primera muestra), 0.54, 0.69, 0.48, 0.62, 0.65, 0.58, etc.

Los matemáticos han mostrado que la probabilidad de una proporción en una muestra al azar cayendo en un intervalo está dada por una curva de probabilidad normal (Figura 4.48) con una media

$$\mu = P$$

(lo cual, desafortunadamente, es desconocido) y una desviación estándar

$$\sigma = \sqrt{\frac{P(1-P)}{n}},$$

donde n es la cantidad de personas en una muestra.

Figura 4.48

En nuestro ejemplo, $n = 1{,}000$, de manera que

$$\sigma = \sqrt{\frac{P(1-P)}{1000}}$$

Sin embargo, aún no conocemos el valor de P. Para muestras *grandes* (1,000 personas es considerada una muestra "grande"), los estadísticos han mostrado que reemplazando P en (1) arriba con la proporción en nuestra muestra no produce error significativo. Por consiguiente, podemos tomar

$\sigma = \sqrt{\frac{.60(.40)}{1000}} = 0.0155$ (llevado a cuatro lugares decimales).

Conocemos, de la sección anterior, que la probabilidad de una proporción sea entre $\mu - 1.96\sigma$ y $\mu + 1.96\sigma$ o, en este caso $P - 1.96\sigma$ y $P + 1.96\sigma$ es 0.95 (Figura 4.49).

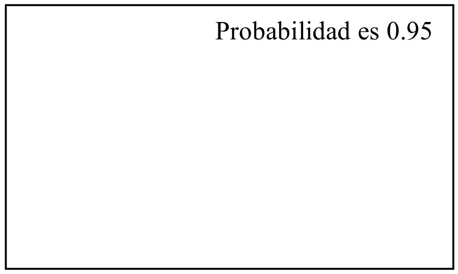

Figura 4.49

Ahora, toma una muestra y computa la proporción en la muestra, y llámala p. En nuestro ejemplo, $p = 0.60$. Forma un intervalo

$$\text{de } p - 1.96\sigma \text{ a } p + 1.96\sigma$$

o, en nuestro ejemplo,

$$\text{de } 0.60 - 1.96(0.0155) \text{ a } 0.60 + 1.96(0.0155).$$

Este intervalo contendrá la proporción de la población P, sí y sólo sí, p está en el intervalo

$$\text{de } P - 1.96 \text{ a } P + 1.96$$

(Figura 4.50). La probabilidad de que esto ocurra es 0.95.

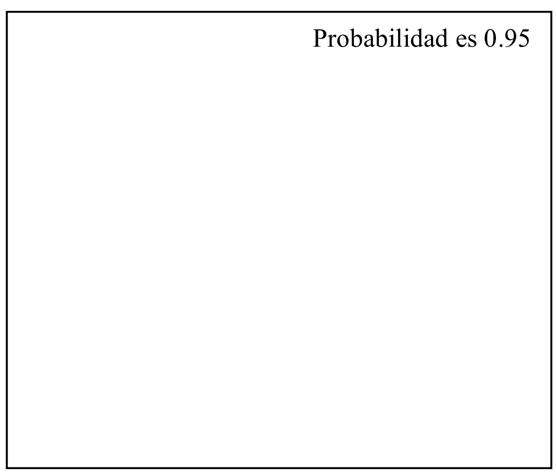

Figura 4.50

Nota que el intervalo en la izquierda incluye el número desconocido P mientras el intervalo en la derecha no lo incluye. En nuestro ejemplo, construimos el intervalo

$$\text{de } 0.60 - 1.96(0.0155) \text{ a } 0.60 + 1.96(0.0155)$$
$$\text{ó}$$
$$\text{de } 0.60 - 0.03 \text{ a } 0.60 + 0.03$$

No sabemos qué tipo de intervalo es este. ¿Incluye u omite este intervalo la proporción actual P de los votantes que apoyan al candidato demócrata? Sin embargo, sabemos que este proceso llevará a intervalos que incluyen P 95% del tiempo. Por esta razón, un intervalo como

$$\text{de } 0.60 - 0.03 \text{ a } 0.60 + 0.03$$

o, en general,

$$\text{de } p - 1.96\sigma \text{ a } p + 1.96\sigma$$

se le conoce como un 95% *intervalo de confianza*.

Volvamos a nuestro problema original: una estación de televisión acaba de llevar a cabo una encuesta entre 1,000 votantes y encontró que 600 escogieron al candidato demócrata sobre el candidato republicano. La encuesta *sugiere* que el candidato demócrata cuenta con el apoyo del 60% de los votantes. Hemos acordado que la proporción actual de los votantes que apoyan al candidato demócrata está dado por

$$0.60 \pm \text{“algún error”}$$

Nuestro trabajo sugiere que tomemos “algún error” para que sea 1.96σ, o, en nuestro ejemplo, 0.03. Esto es, con 95% de confianza (usando el significado de arriba), podemos decir que la proporción actual de votantes que apoyan el candidato demócrata se encuentra entre

$$0.60 - 0.03 \text{ y } 0.60 + 0.03$$

o, entre

$$0.57 \text{ y } 0.63$$

El número 0.03 ó 3% es conocido algunas veces como el *margen de error*.

Si la estación de televisión llevara a cabo una encuesta de una muestra al azar diferente de 1,000 votantes, podría obtener un 95% de intervalo de confianza tal como entre

0.54 y 0.60

Este intervalo es diferente del anterior. La idea es que si ellos hicieran esto muchas veces, el 95% del tiempo ellos serían exitosos en tener intervalos que incluyeron la proporción actual de los votantes que apoyan al candidato demócrata.

Para que puedas obtener alguna práctica para encontrar 95% de intervalos de confianza, trata los siguientes dos problemas:

1. **Una encuesta del National Post de 1,400 votantes indicó que 950 apoyaron al candidato demócrata. Encuentra un 95% de intervalo de confianza para la proporción de la población que apoya al candidato republicano.**

2. **Una encuesta de 1,200 votantes llevada a cabo por la estación radial WMC encontró que 750 apoyaron al candidato republicano. Encuentra un 95% de intervalo de confianza para la proporción de la población que apoya al candidato republicano.**

En nuestro ejemplo anterior, se llevó a cabo una encuesta de 1,000 personas para determinar las proporciones de votos para los candidatos en cierto estado. En esta muestra, 600 personas apoyaron al candidato demócrata, lo que llevó a una *sugerencia* de que el 60% de la población apoyó este candidato. Si la encuesta (sondeo) fue tomada en una región demócrata del estado, es poco probable que uno pueda atribuir mucho significado a esta sugerencia. Cuando se lleva a cabo cualquier encuesta, la primera pregunta debe ser: "¿Cómo y dónde se llevó a cabo la encuesta?" Si la muestra no es una representación verdadera de la población, entonces, es muy probable que ocurran errores en las inferencias hechas de la muestra.

Recuerda, la *población* contiene el conjunto completo de los individuos para las cuales las inferencias de la muestra se van a hacer. Permite que N sea la cantidad total de elementos en el conjunto que compone la población. Permite que n sea la cantidad de elementos en una muestra. Los elementos n de una muestra deben ser seleccionados de manera que la muestra sea representativa de la población. Cada elemento en la población debe tener una posibilidad igual (no cero) de ser seleccionado en la muestra. Por ejemplo, es un error tomar nombres de una guía telefónica, ya que algunas personas no tienen teléfono o sus números telefónicos no están listados en la guía y no están representados propiamente en la muestra.

Llevar a cabo una encuesta con personas en la calle es a menudo parcial porque el entrevistador puede seleccionar personas que se ven bien y están bien vestidas. Un congresista o una congresista no pueden depender de la correspondencia como una muestra imparcial de las personas representadas, ya que esto es una muestra de personas con opiniones fuertes e incluye una gran cantidad de personas raras y miembros de interés especial.

PHOTO

Otra fuente de *prejuicio* es la localización donde la encuesta es llevada a cabo. Por ejemplo, una encuesta de las primeras mil personas encontradas en una esquina de una calle en Nueva Orleáns no sería considerada una muestra al azar de la población de los Estados Unidos. Es necesario incluir personas del oeste, del este, y así sucesivamente.

Enumera por lo menos dos otras dificultades que veas en seleccionar una muestra de una población.

Un método ideal para escoger una muestra es llamado Muestreo Simple al Azar (SRS, pos sus siglas en inglés). Para poder usar SRS cada uno de los N elementos en la población necesita que se le asigne un número de 1 a N. Una vez los elementos de la población han sido numerados, selecciona una muestra (subconjunto) n de estos números entre 1 y N por un proceso al azar. Esto significa, por ejemplo, usar números al azar de tu TI-84 Plus (TI-83 Plus) o de un número de una tabla al azar (recuerda anteriormente en **MATH** *Connections*).

Ya que cada elemento tiene una oportunidad igual de ser seleccionado y hay

$$\binom{N}{n}$$

muestras posibles, cada muestra tiene una probabilidad de

$$\frac{1}{\binom{N}{n}}$$

de ser seleccionada.

El Muestreo simple al azar provee un estimado *imparcial* de las características de la población, tales como las proporciones que apoyan los asuntos o candidatos. Donde hay una diferencia significativa entre la población y la muestra, hay **prejuicio**. Por ejemplo, si solamente tomaste muestras de los miembros de los Republicanos Jóvenes para determinar por quiénes ellos votarían en la próxima elección y de este resultado implicar que la población general votaría de la misma manera, tendrías un estimado *prejuiciado*.

Describe la población que usarías para cada una de las siguientes. En cada caso, ¿cómo seleccionarías una muestra para asegurarte que el prejuicio no causa problemas con tus estimados (conclusiones)?

1. **Se condujo un estudio para determinar si el juego de pelota era más popular que el juego de fútbol.**

2. **La Junta Estatal de Educación quería determinar si la puntuación en las pruebas SAT eran más altas en las escuelas urbanas que en las escuelas rurales.**

3. **El gobierno de los Estados Unidos quiere determinar la cantidad de individuos que viven en los Estados Unidos.**

Terminamos esta sección con una observación final sobre las muestras. Un problema que no es mencionado aquí es el tamaño de una muestra. ¿Cuán grande debe ser el tamaño de la prueba que vas a tomar? Esto es, ¿cuál debe ser el tamaño de n? Intuitivamente, mientras más grande es N, más grande debe ser n. No se debe esperar que una muestra de 2 de una población de 1,000 te lleve a un muy buen estimado, no importa cuán al azar estas 2 personas hayan sido seleccionadas. Cuando continúes tu estudio de la estadística, cómo seleccionar un tamaño apropiado de la muestra es uno de los problemas que aprenderás a resolver.

REFLEXIONA

En este capítulo, hemos cubierto una gran cantidad de ideas las cuales te deben capacitar para resolver muchos problemas significativos. Fuiste de contar herramientas, tales como las permutaciones y las combinaciones, a herramientas de probabilidad, tales como los eventos independientes y distribuciones binómicas. Más adelante, se te mostró como la probabilidad es un modelo básico para un estudio de estadísticas. Se espera que percibas la incertidumbre en el mundo a tu alrededor, pero, que realices que hay herramientas que le permiten a uno comprender y usar esta incertidumbre de maneras productivas.

Conjunto de ejercicios: 4.9

1. Las experiencias pasadas indican que cerca del 25% de las personas que contraen cierto tipo de enfermedad mueren de dicha enfermedad. Por consiguiente, las personas en el campo de la medicina asumen que la probabilidad de muerte por esta enfermedad es $\frac{1}{4}$. Una droga medicinal nueva es administrada a 40 personas que padecen la enfermedad y solamente 1 persona muere de dicha enfermedad. ¿Debe la droga nueva ser declarada como efectiva? Explica.

2. La estación radial WMC llevó a cabo una encuesta entre 800 votantes y encontró que 523 apoyaron al candidato demócrata. Encuentra un intervalo de confianza de 95% para la proporción de la población que apoya al candidato demócrata.

3. Una encuesta del periódico Republicano-Americano de 1,500 votantes encontró que 924 apoyan al candidato republicano. Encuentra un intervalo de confianza de 95% para la proporción de la población que apoya al candidato republicano.

4. En un ejemplo dado en esta sección, una encuesta de 1,000 personas encontró 600 personas apoyando al candidato demócrata. El proceso de construir un intervalo de confianza de 95% fue descrito. Para este mismo ejemplo, describe el proceso de construcción de un intervalo de confianza de 99%.