equipo de ingeniería de proyectos utilizó unidades métricas de medida (es decir, partes de metros) para comunicarse con otro equipo que supuso que los números estaban en unidades inglesas (partes de pulgadas). Esta confusión acerca de las unidades de medida costó a la NASA $125 millones de dólares (http://www.cnn.com/TECH/space/9909/30/mars.metric/).

Codificación y conteo de observaciones

Una vez completa la recolección de datos, el siguiente paso en el manejo de datos consiste en codificar y registrar todas las mediciones en una hoja de cálculo o en un archivo de datos de computadora. La tabla 2-4 presenta un ejemplo de un **registro o guía de codificación**, que es *una descripción concisa de símbolos que describen el significado de cada puntuación para cada variable.* Tales datos vienen de un estudio ficticio sobre estudiantes del Instituto Apple Pond. En este registro de codificación sustituimos símbolos numéricos por las categorías de masculino y femenino y por niveles de escolaridad. Esto se hace porque a las computadoras se les facilita contar y seleccionar números (en lenguaje computacional, símbolos *numéricos*) que palabras (*caracteres* o símbolos de *cadena*).

En un registro de codificación, se debe tener cuidado en ser muy preciso porque la codificación de respuestas podría introducir un error de medición. Cada variable se codifica siguiendo dos principios básicos: inclusividad y exclusividad. El **principio de inclusividad** establece que para una variable determinada *debe haber una puntuación o un código para cada observación realizada.* Dicho de otra manera, ¿incluimos una categoría de respuesta o puntuación para toda respuesta posible? Por ejemplo, con la variable nominal *raza* podríamos indicar las categorías de blanco, afroamericano, nativo americano, asiático-americano, hispano y otra(s). La categoría de respuesta *otra(s)* evita la necesidad de ocupar espacio en el cuestionario para las categorías que se espera tengan pocas respuestas en el lugar del estudio (esquimales en Kansas). El código otra(s) es una *categoría residual* que abarca los remanentes (piensa en la palabra *residuo*).

Aun cuando ignoramos la cuestión de cómo codificar a las personas de ascendencia mixta, si sólo usamos blanco y afroamericano, la categoría de raza no será inclusiva de, por decir, asiático-americano. Sin su propia categoría u otra(s) no es factible suponer que todos los asiático-americanos se registrarán de la misma forma. Algunos anotarán blanco, pero otros quizá dejen sin contestar el espacio de la pregunta. Después de calcular los totales quizá no

TABLA 2-4 | Registro de codificación para respuestas de cuestionarios de especialistas en justicia en el Instituto Apple Pond

Nombre de la variable	Descripción de códigos de la variable
NOMBRE DEL ESTUDIANTE	Registre el nombre, apellidos paterno y materno; espacio en blanco = faltante
EDAD	Registre la edad informada hasta 97 años; 98 = 98 años o más; 99 = faltante
GÉNERO	0 = hombre, 1 = mujer, 9 = faltante
PROMEDIO	Promedio en una escala de cuatro puntos = número de puntos de calidad ganados por hora crédito (redondeado a dos lugares decimales); 9.99 = faltante
ESCOLARIDAD	1 = nuevo ingreso; 2 = segundo año; 3 = intermedio inicial; 4 = intermedio avanzado; 8 = otro; 9 = faltante

seamos capaces de señalar exactamente cuántos tenemos de *cualquier* raza. Perdimos a los asiático-americanos que no contestaron la pregunta, y algunos de nuestros "blancos" son asiático-americanos pero no tenemos noción de *cuántos.* Semejante descuido de pérdida de control sobre el error de medición puede hacer que los datos de raza sean inservibles.

El **principio de exclusividad** sostiene que para una variable determinada *a cada observación se asigna una y sólo una puntuación.* Así, cada categoría debe excluir puntuaciones que no le pertenezcan y cualquiera de las dos categorías no debe compartir una respuesta. Por ejemplo, con la variable *afiliación religiosa en la niñez,* las categorías de respuesta protestante, bautista, católico, judío y otra(s) no serían mutuamente excluyentes porque un bautista quizá se anote como bautista, mientras otro lo haga como protestante. Cuando se sumen los totales de cada categoría, no podremos decir cuántos bautistas había en la muestra. Algunos quizá se registraron como "protestante", pero no tenemos forma de especificar quiénes y cuántos lo hicieron. La tabla 2-5 muestra los resultados del Estudio Social General de 1994 para esta variable. La exclusividad se asegura formulando a los protestantes las preguntas adicionales necesarias para conocer sus denominaciones específicas. (Para ayudar a la comprensión sobre la información de las tablas de este texto, modificaremos las tablas para diferenciar claramente las "Especificaciones" de los datos disponibles y los "Cálculos". En los reportes publicados de estas tablas, dichos términos no aparecerían.)

TABLA 2-5 | Distribución de la afiliación religiosa en la niñez, respuesta a las preguntas: ¿en qué religión fue educado? Si fue protestante, ¿en qué denominación específica, si la hay?

Especificaciones		Cálculos
Categoría de respuesta	*a)* Número	*b)* Porcentaje de la muestra total
Protestante		
Bautista	706	23.73
Metodista	319	10.72
Luterano	220	7.39
Presbiteriano	139	4.67
Episcopal	68	2.29
Otra	309	10.39
Ninguna denominación o iglesia sin denominación	69	2.32
Total protestante	1 830	61.51
Católica	882	29.65
Judía	55	1.85
Ninguna	127	4.27
Otra	74	2.49
Sin respuesta	7	0.23
Total	2 975	100.00

Fuente: National Opinion Research Center, General Social Survey 1994.
www.icpsr.umich.edu/gss/codebook/relig16.htm
www.icpsr.umich.edu/gss/codebook/denom16.htm

TABLA 2-6 | Hoja de cálculo de respuestas al cuestionario de 10 especialistas en justicia delictiva en el Instituto Apple Pond (datos ficticios)

	Especificaciones			
Nombre del estudiante	**Edad**	**Género**	**Promedio**	**Escolaridad**
Jessica A Cortland	19	1	3.21	2
Mark E Pippin	22	0	2.75	4
Stayman V Winesap	19	0	2.43	1
Barry D McIntosh	21	0	3.39	3
Harriet G Smith	20	1	3.87	3
Antonio B Rome	22	0	2.32	3
Robert J Cox	18	0	3.25	1
Rodney I Greening	20	0	9.99	2
Thomas R York	22	0	2.47	4
Goldie D Licious	19	1	3.68	2

Regresa al registro de codificación de la tabla 2-4 y observa que el principio de inclusividad se cumple proporcionando un *código para los datos perdidos*, llamados **valores perdidos**. Decimos, por ejemplo, que el género y escolaridad tienen un valor perdido de 9. En algunos estudios, los valores perdidos se presentan cuando por accidente el entrevistador se salta una pregunta o un encuestado no contesta. Al calcular los estadísticos para una variable, pasamos por alto los casos que resulten en un valor perdido.

La tabla 2-6 es una hoja de cálculo de los resultados de la aplicación del cuestionario empleado en una encuesta aplicada a 10 especialistas en justicia delictiva del Instituto Apple Pond. Una hoja de cálculo es una matriz que muestra las puntuaciones de todas las variables organizadas en columnas; y todos los casos, en filas.

Una hoja de cálculo es útil para ordenar y resumir datos de una forma eficaz. Por ejemplo, si contamos rápidamente el número de mujeres en la muestra sumando las unidades citadas bajo "Género". Mediante esta simple hoja de cálculo podemos ver rápidamente que la muestra se compone de siete hombres y tres mujeres; existen dos estudiantes de primer año, tres de segundo año, tres de tercer año y dos de último año; el rango de edades oscila entre los 18 y los 22 años, y el rango del promedio va de 2.43 a 3.87 con un caso no reportado. Por supuesto, para una muestra grande, un procedimiento eficaz implica tanto la especificación de estos códigos de la hoja de cálculo en un archivo de datos de la computadora como hacer que el programa computacional se encargue de los cálculos. Los archivos de datos de computadora están organizados como estas hojas de trabajo.

Distribuciones de frecuencias

Una vez que todos los datos están organizados en una hoja de cálculo o en un archivo de datos de computadora, el siguiente paso en el análisis consiste en enfocarse por separado en cada variable y contestar la pregunta: ¿cuántos sujetos caen en cada categoría o puntuación? Organizamos los datos de cada variable en una **distribución de frecuencias**, que es *una lista de todas las puntuaciones observadas de una variable y la frecuencia (f) de cada puntuación (o categoría)*. Utilizamos letras mayúsculas para representar una variable. Si X se define como la variable *género*, la distribución de frecuencias de X simplemente muestra cuántos hombres y mujeres hay en la muestra. La tabla 2-5 mostrada anteriormente proporciona la distribución de frecuencias para la afiliación religiosa en la niñez.

Distribución de frecuencias Lista de todas las puntuaciones observadas de una variable y la frecuencia (f) de cada puntuación (o categoría).

Estandarización de distribuciones de puntuaciones

El conocimiento de la frecuencia de una categoría no resulta muy informativo por sí mismo. Por ejemplo, alguien nota que existen cinco millonarios que viven en una ciudad. Cinco no son muchos para la ciudad de Nueva York, pero lo serían para un pueblo de 800 personas. Así, es más informativo reportar la frecuencia de una categoría como una proporción o porcentaje con respecto al número total de sujetos de la muestra. La imaginación estadística nos impulsa a expresar la frecuencia de una categoría en un contexto mayor, como una parte en relación con un todo. Planteamos la pregunta: cinco millonarios ¿de cuántas personas? Como observamos en el capítulo 1, las fracciones, las proporciones y los porcentajes ofrecen denominadores comunes o "medidas estándar" para facilitar la comparación de categorías y muestras. Para una muestra como un todo, la **distribución de frecuencias con proporciones** consiste en *una lista de la proporción de respuestas para cada categoría o puntuación de una variable.* La **distribución de frecuencias de distribución** es *una lista del porcentaje de respuestas para cada categoría o puntuación de una variable.*

Distribución de frecuencias con proporciones Lista de la proporción de respuestas para cada categoría o puntuación de una variable.

Distribución de frecuencias de porcentajes Lista del porcentaje de respuestas para cada categoría o puntuación de una variable.

Para obtener estas distribuciones para cada categoría de respuesta o puntuación de una variable, escribimos una fracción y después la dividimos para obtener la proporción y el porcentaje. Para facilitar la interpretación, la distribución de frecuencias de porcentajes es la que comúnmente se reporta. Por ejemplo, en los datos de la hoja de cálculo del Instituto Apple Pond de la tabla 2-6 podemos observar que el porcentaje de hombres es

$$p \text{ [de la muestra de estudiantes que son hombres]} = \frac{\#\text{ hombres}}{n} = \frac{7}{10} = 0.7000$$

$$\% \text{ [de la muestra de estudiantes que son hombres]} = (p)\,(100) = (0.7000)\,(100) = 70.00\%$$

donde p es la proporción y n es el tamaño de la muestra. Después de hacer lo mismo para las mujeres, tenemos la distribución de frecuencias de los porcentajes de la variable *género* en la columna de la derecha de la tabla 2-7. En ésta también se incluyen la frecuencia y las distribuciones de frecuencias proporcionales. El total de todas las proporciones y porcentajes para la distribución será igual a 1.0000 y 100.00 por ciento, respectivamente, considerando el error de redondeo.

TABLA 2-7 | Frecuencia, frecuencia proporcional y distribuciones de frecuencias porcentuales de la variable género para una muestra de 10 estudiantes del Instituto Apple Pond

Especificaciones		Cálculos	
Género (X)	Frecuencia (f)	Frecuencia proporcional	Frecuencia porcentual (%)
Hombre	7	0.7000	70.00
Mujer	3	0.3000	30.00
Total	10	1.0000	100.00

Cálculo de las frecuencias proporcionales y porcentuales de una categoría

$$p\ [\text{de la muestra total } (n) \text{ en una categoría}] = \frac{f \text{ de una categoría}}{n} = \frac{\#\text{ en categoría}}{n}$$

$$\%\ [\text{de la muestra total } (n) \text{ en una categoría}) = (p\ [\text{de la muestra total } (n) \text{ en una categoría}])\ (100)$$

donde

p [de la muestra total (n) en una categoría] = proporción de todos los casos que caen en la categoría,

f = frecuencia de casos (o número de casos) en la categoría,

n = tamaño de la muestra

La tabla 2-5 (en la página 49) muestra la frecuencia y la distribución de frecuencias con porcentajes para la variable afiliación religiosa en la niñez.

Codificación y conteo de datos de intervalo/razón

Las variables con niveles de medición de intervalo/razón se distinguen de las variables nominales/ordinales por sus cualidades numéricas, sobre todo por sus unidades de medición, como millas, kilómetros, pulgadas, segundos y kilogramos. Tales variables "cuantitativas" nos permiten imaginar una regla y pensar linealmente en términos de la distancia entre puntos sobre una línea recta. Es más, podemos hacer mediciones muy precisas.

Una **medición precisa** es *aquella en la que el grado de error de medición es suficientemente pequeño para la tarea en cuestión.* La precisión depende de circunstancias prácticas y se controla al especificar el error de redondeo. Por ejemplo, al cortar dos troncos de dos pies de largo para una chimenea, un pequeño grado de precisión será suficiente porque podemos

darnos el lujo de tener un elevado grado de error, por ejemplo, "medio pie de más o de menos". En contraste, para una prueba de calidad de tarjetas de circuitos de microcomputadoras puede exigirse una precisión de una milésima de pulgada. El grado de precisión es cuestión de tolerancia. Preguntamos: ¿qué error de medición podemos tolerar (o soportar) sin encontrar problemas prácticos o sacar conclusiones científicas con falla?

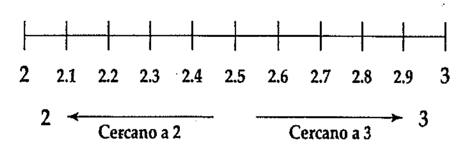
Redondeo de las observaciones de intervalo/razón

La observación de una variable de intervalo/razón tal vez no nos ofrezca la puntuación verdadera porque sus mediciones a menudo pueden hacerse de manera infinitamente más precisa. Por ejemplo, podemos medir distancias hasta la unidad más cercana en kilómetros, metros, centímetros, y así sucesivamente. Por tanto, redondeamos las puntuaciones de intervalo/razón hasta cierto grado de precisión elegido y especificado. De esta manera, reconocemos que el código registrado para la puntuación tiene algún error de medición. El **error de redondeo** es *la diferencia entre la puntuación real o perfecta (que quizá nunca conozcamos) y nuestra puntuación observada y redondeada.* El error de redondeo depende de qué *posición decimal elegimos como nuestro nivel de precisión,* nuestra unidad de redondeo. (Si es necesario, el estudiante debe revisar en el apéndice A la ubicación de las posiciones decimales.) Si decidimos medir el tiempo a la centésima de segundo más cercana, como se efectúa en los eventos olímpicos de pista, entonces nuestra unidad de redondeo es la posición de las centésimas.

El procedimiento para redondear una puntuación de una variable de intervalo/razón es como sigue:

1. Especifica la unidad de redondeo según su posición decimal.
2. Observa el número *a la derecha* de la unidad de redondeo y sigue estas reglas:
   A. Si es 0, 1, 2, 3 o 4, redondea hacia el entero inferior.
   B. Si es 6, 7, 8 o 9, redondea hacia el entero superior.
   C. Si es 5, observa la siguiente posición decimal a la derecha y, si el número es 5 o mayor, redondea hacia el entero superior; si no existe algún número en esa siguiente posición decimal, deja el redondeo en ese número.

Piensa en el redondeo como un movimiento hacia el punto más cercano sobre una línea. Por ejemplo, si redondeamos al *entero más cercano* (la posición de las unidades), simplemente estamos moviendo al entero más cercano.



Por tanto, aquí 2.1, 2.2, 2.3 y 2.4 se redondean hacia abajo a 2 sólo porque están más cerca de 2 que de 3. Se ofrecen ejemplos adicionales en el apéndice A.

Los límites reales de puntuaciones redondeadas

Una vez que conocemos las puntuaciones redondeadas, los números *en la posición decimal de la unidad de redondeo* se consideran estimaciones. El valor real de una puntuación podría ser cualquiera de las puntuaciones que se redondean para obtener la puntuación registrada.

Por ejemplo, supongamos que redondeamos la estatura de Jonathan a la pulgada más cercana y registramos 69 pulgadas. Varias horas después, con Jonathan ausente, Alan nos pregunta cuál es la estatura de Jonathan. Observamos nuestra hoja de cálculo de datos y vemos el código de 69 pulgadas. En este momento, podemos afirmar que la estatura real de Jonathan está entre 68½ y 69½ pulgadas, o 69 más menos media pulgada. Este *rango de posibles valores reales de una puntuación (ya) redondeada* se llama **límite real** o **límite verdadero** de la puntuación.

Los límites reales de una puntuación redondeada especifican el rango de números que podrían redondearse para obtener la puntuación registrada. En este sentido, calcular los límites reales es la inversa del redondeo. Por ejemplo, supongamos que registramos cuánto tiempo toma a cada uno de los 150 estudiantes completar un proyecto de laboratorio de química y lo redondeamos a la hora más cercana. También considera que 56 de estos estudiantes reciben una puntuación de dos horas. Algunos de ellos tomaron un poco menos de dos horas y otros un poco más. Precisamente, un estudiante que registra dos horas pudo haber tomado sólo 90 minutos (1½ horas), el *límite real inferior*, o hasta 150 minutos (2½ horas), el *límite real superior.* Una puntuación de dos horas *en realidad* significa entre 1½ y 2½ horas; por eso llamamos límites reales a este rango de tiempos.

Límite real inferior de dos horas 1½ (90 minutos)	← 2 horas (120 minutos) →	Límite real superior de dos horas 2½ (150 minutos)

Calculamos los límites reales moviéndonos media unidad de redondeo en cada dirección, utilizando el siguiente procedimiento:

Cálculo de límites reales de una puntuación de intervalo/razón

1. Observa la puntuación e identifica la "unidad de redondeo", el lugar decimal al que la puntuación se redondeó (como en la columna B que sigue). (Para ubicaciones del lugar decimal, revisa la figura A-1 del apéndice A.)
2. Divide entre 2 esta unidad de redondeo (como en la columna C que sigue). *Atención:* no dividas el número del lugar decimal de la unidad de redondeo entre 2.
3. Resta el número del paso 2 de la puntuación redondeada observada, para obtener el límite real inferior (LRI, como en la columna D que sigue).
4. Suma el resultado del paso 2 a la puntuación redondeada observada, para obtener el límite real superior (LRI, como en la columna E que sigue).

Ejemplos:

	(A) Calcula los límites reales de:	(B) Identifica la unidad de redondeo	(C) Divide entre 2 la unidad de redondeo	(D) LRI Resta (C) de (A)	(E) LRS Suma (C) a (A)
a)	.48	.01	$\frac{.01}{2} = .005$	.475	.485
b)	17	1	$\frac{1}{2} = .5$	16.5	17.5
c)	4000	1000	$\frac{1000}{2} = 500$	3500	4500
d)	0.7	.1	$\frac{.1}{2} = .05$	.65	.75

Por ejemplo, para los 56 estudiantes que calificaron dos horas en el proyecto del laboratorio de química, redondeamos a la hora más cercana (la posición de las unidades). Dividimos *esta unidad de redondeo de una hora* entre 2 para obtener hora y media. Entonces restamos este resultado de la puntuación redondeada observada de dos horas para obtener el límite real inferior (1½ horas) y lo sumamos a la puntuación observada de dos horas, para obtener el límite real superior (2½ horas). Incluso es improbable que uno de estos 56 estudiantes tomara exactamente dos horas para completar el proyecto; dos horas es una estimación redondeada. Podemos tener la certeza, sin embargo, de que cada uno de los 56 terminara entre 1½ y 2½ horas. Nuestro grado de precisión es la unidad de redondeo de una hora.

Los principios de inclusividad y exclusividad también se aplican a las variables de intervalo/razón. Para una variable como la edad, apegarse al principio de inclusividad parecería razonable; sólo registramos la "edad en el último cumpleaños". No obstante, para garantizar la inclusividad, un cuestionario de investigación debe incluir las respuestas "se negó" y "no sabe". La exclusividad es razonable en cuanto a que todas las mediciones se realicen de la misma manera, en este caso la edad en el último cumpleaños. Si un encuestado dice que tiene 26 años, entonces registra 26, no 27 ni 25.

Distribuciones de frecuencias de proporciones y de porcentajes para variables de intervalo/razón

Las distribuciones de frecuencias de proporciones y porcentajes para variables de intervalo/razón se calculan de la misma forma que para variables nominales/ordinales, excepto que en lugar de categorías tenemos puntuaciones. Por ejemplo, si la Universidad Smithville tiene 10 000 estudiantes y 3 000 tienen 19 años, las frecuencias proporcionales y porcentuales para la puntuación de 19 años son

$$p\,[\text{de 19 años en la Universidad Smithville}] = \frac{f\text{ de 19 años}}{n} = \frac{3\,000}{10\,000} = 0.3000$$

$$\%\,[\text{de 19 años en la Universidad Smithville}] = (p)\,(100) = 30.00\%$$

TABLA 2-8 | Ilustración de una distribución de frecuencias porcentuales acumuladas: años de escolaridad entre cuidadores de pacientes ancianos con Alzheimer

Especificaciones		Cálculos	
Años de educación formal (X)	Frecuencia (f)	Frecuencia porcentual	Porcentaje acumulado (%) (f)
5	1	5%	5
6	1	5	10
7	1	5	15
9	2	10	25
10	1	5	30
11	1	5	35
12	10	50	85
14	2	10	95
16	1	5	100
Total	20	100%	100

Si estos cálculos se realizan para todas las edades, los resultados se presentan como la distribución de frecuencias de porcentaje de la variable *edad* para la población de estudiantes de la Universidad de Smithville.

Distribuciones de frecuencias de porcentajes acumulados

La tabla 2-8 presenta la frecuencia, la frecuencia de porcentaje y las distribuciones de frecuencias de porcentajes acumulados de los niveles de escolaridad de 20 cuidadores, parientes que acompañan a pacientes con Alzheimer en una clínica (Clair, Ritchey y Allman, 1993). Estas tres piezas de información son partes típicas de los resultados obtenidos por computadora porque juntos generan respuestas rápidas a una serie de preguntas. Obviamente, la frecuencia de puntuación bruta (*f*) proporciona una respuesta sobre cuántos sujetos recibieron una puntuación específica, y la frecuencia porcentual estandariza la frecuencia de acuerdo con el tamaño de la muestra. La información adicional de la tabla 2-8, la frecuencia de porcentajes acumulados, es una valiosa forma para observar las frecuencias de las puntuaciones en una distribución hasta, e inclusive, una puntuación de interés. Ésta es la **frecuencia de porcentajes acumulados**, que es *la frecuencia porcentual de una puntuación y además la de todas las puntuaciones que la preceden en la distribución.* Por ejemplo, en el caso de los cuidadores de la tabla 2-8, ¿qué porcentaje tiene un nivel de escolaridad hasta e inclusive el nivel de preparatoria? Para obtener las frecuencias de porcentajes acumulados hacemos una lista con las puntuaciones, de la más baja a la más alta, y calculamos la frecuencia de porcentaje de cada puntuación. Entonces sumamos las frecuencias de porcentaje de la puntuación que nos interesa y todas las puntuaciones menores. En la tabla 2-8, 85 por ciento tenían 12 años de escolaridad o menos. Restando esta frecuencia de porcentajes acumulados del 100 por ciento, rápidamente podemos ver que sólo 15 por ciento de la muestra fueron más allá de la escuela preparatoria.

El siguiente cuadro es una guía sobre cómo elaborar distribuciones de frecuencias.

Para elaborar distribuciones de frecuencias

Supongamos que deseamos elaborar una tabla de distribución de frecuencias para la variable de edad de la tabla 2-6, para una muestra de ficción del Apple Pond Institute. La tabla 2-6 presenta los datos en formato de hoja de cálculo. Debemos completar la siguiente tabla A para presentar la distribución de frecuencias, distribución de frecuencias porcentuales y distribución de frecuencias de porcentajes acumulados para la variable *edad*. Sigue estos pasos:

1. Construye una plantilla con el título y encabezados apropiados (la información arriba de las columnas de números en la tabla A).
2. Observa las puntuaciones para edades entre los 10 estudiantes de la tabla 2-6 y haz una lista, sólo una vez, de cada uno de los valores de X del más bajo al más alto. Esto va bajo "Edad (X)" en la tabla A.
3. Cuenta el número de estudiantes para cada edad de la tabla 2-6 e inserta esta cantidad bajo "Frecuencia (f)" en la tabla A. Comprueba ver que la frecuencia total es el tamaño muestral, $n = 10$, y registra este total.
4. Calcula la frecuencia proporcional para cada valor de X al dividir cada frecuencia entre el total n de 10. Estos resultados van bajo "Frecuencia porcentual" en la tabla A. Comprueba que el total de la frecuencia porcentual suma 1.0000. Si los cálculos son correctos y el total no suma 1.0000, entonces inserta una nota al pie que indique "El total no sumó 1.0000 por error de redondeo".
5. Calcula la frecuencia de porcentaje de cada uno de los valores de X al multiplicar por 100 la frecuencia proporcional (es decir, al mover el punto decimal dos lugares a la derecha). Estos resultados van bajo "Frecuencia porcentual" en la tabla A. Si los cálculos son correctos y el total no suma 100.00%, entonces inserta una nota al pie que indique "El total no sumó 100 por ciento por error de redondeo".
6. Para la tabla A, calcula la frecuencia de porcentajes acumulados, la frecuencia de porcentaje de una puntuación más la de todas las puntuaciones que la preceden. Empieza por registrar 10.00 por ciento para $X = 18$. Ahora suma esta frecuencia de porcentajes acumulados de $X = 18$ al *porcentaje* de frecuencia de $X = 19$ para obtener el porcentaje de frecuencia acumulada de $X = 19$, que es 10.00 por ciento + 30.00 por ciento = 40.00 por ciento. Ahora suma esta frecuencia de porcentajes acumulados de $X = 19$ a la frecuencia *porcentual* de $X = 20$ para obtener la frecuencia de porcentajes acumulados de $X = 20$, que es 40.00 por ciento + 20.00 por ciento = 60.00 por ciento, y así sucesivamente. Asegúrate de que la frecuencia de porcentajes acumulados del valor más alto de X sume 100.00 por ciento.

TABLA A. | Distribuciones de frecuencia, de frecuencia proporcional y de frecuencia porcentual de la variable *edad* de 10 estudiantes del Apple Pond Institute

Edad (X)	Frecuencia (f)	Frecuencia proporcional	Frecuencia porcentual (%)	Frecuencia porcentual acumulativa (%)
18	1	.1000	10.00	10.00
19	3	.3000	30.00	40.00
20	2	.2000	20.00	60.00
21	1	.1000	10.00	70.00
22	3	.3000	30.00	100.00
Totales	10	1.0000	100.00	

Percentiles y cuartiles

Con frecuencia visualizamos una distribución de puntuaciones como fraccionada o "fracturada" en grupos que están arriba y debajo de una puntuación, o en grupos con iguales porcentajes de casos. Las distribuciones de frecuencias acumuladas proporcionan una herramienta para identificar **cuantiles,** *puntuaciones que separan una fracción de los casos de una distribución.* Los rangos percentilares (o simplemente percentiles) son un cuantil común. *Entre los casos en una distribución de puntuaciones,* el **rango percentilar** es *el porcentaje de casos que caen en o están debajo de un valor específico de X.* Por ejemplo, en las frecuencias de porcentajes acumulados de la tabla 2-8, vemos que un cuidador con 14 años de escolaridad tiene un nivel educativo igual o superior al 95 por ciento de la muestra, un rango percentilar de 95. Con frecuencia los percentiles se emplean en círculos de educación como una manera de ordenar notas o calificaciones de exámenes. Por ejemplo, en un examen de admisión a la universidad, un estudiante con una calificación que corresponda al percentil 90 o mayor calificaría para la admisión en una universidad de prestigio, ya que significa que está por encima del 90% del resto de los alumnos.

Pasos para calcular percentiles

El cálculo de un percentil analiza la siguiente pregunta: ¿una calificación particular es igual o más alta que cuál porcentaje de calificaciones? A continuación aparecen datos fraccionales para el examen de un curso. Nótese que las calificaciones están ordenadas en forma ascendente (es decir, de menor a mayor). *Antes de calcular un percentil, las calificaciones deben ordenarse de menor a mayor o de mayor a menor.*

Calculemos el rango percentilar de Taylor en este examen. Su calificación de 78 es igual o más alta que la de 14 de los 27 estudiantes. Primero, calculamos la proporción de casos iguales o menores a 78 y luego el porcentaje.

$$p\,[\text{de calificaciones} \leq 78] = \frac{14}{27} = 0.5185 \qquad \%\,[\text{de calificaciones} \leq 78] = (p)\,(100) = 51.85 = 52\%$$

Entonces, el rango percentilar es 52. Ella alcanzó una calificación igual o mayor a la del 52 por ciento de los estudiantes. Nótese que redondeamos el porcentaje a dos lugares porque los rangos percentilares se reportan en porcentajes enteros (es decir, sin lugares decimales).

Calculemos ahora el rango percentilar de John:

$$p\,[\text{de calificaciones} \leq 91] = \frac{23}{27} = .8518$$

$$\%\,[\text{de calificaciones} \leq 91] = (p)\,(100) = 85.18 = 85\%$$

Entonces, el rango percentilar es 85. Nótese que la calificación de Barry se incluyó en el cálculo porque es igual a la de John.

Lugar de estudiante	Nombre de estudiante	Calificación de examen (ordenado)	Lugar de estudiante	Nombre de estudiante	Calificación de examen (ordenado)
1	Kevin	54	15	Shannel	79
2	Carl	58	16	William	80
3	Robert	61	17	Angie	82
4	Brian	61	18	Akilah	83
5	Maria	65	19	Daniel	85
6	Sean	69	20	Kaitlin	88
7	Jim	70	21	Marcy	90
8	Jessica	72	22	John	91
9	Carol	73	23	Barry	91
10	Brooke	75	24	Wnda	93
11	Kia	75	25	Sarah	95
12	Terry	77	26	Charles	96
13	Jackie	77	27	Elisa	97
14	Taylor	78			

RESUMEN DE PASOS PARA CALCULAR PERCENTILES:

Paso 1. Ordenar las calificaciones.
Paso 2. Calcular la proporción y porcentaje de casos con calificaciones iguales o menores que el caso de interés.
Paso 3. Indicar el percentil en porcentajes enteros.
Nota: Recordar que los percentiles se obtienen fácilmente de una distribución de porcentajes acumulada.

Los **cuartiles** son *cuantiles que identifican las puntuaciones que dividen una distribución en cuatro grupos de igual tamaño (es decir, 25 por ciento de los casos en cada grupo).* Cuando una distribución tiene un rango grande de puntuaciones, los cuartiles se obtienen fácilmente a partir de distribuciones de frecuencias de porcentajes acumulados. El primer cuartil, Q_1, es el 25o. percentil; el segundo, Q_2, es el 50o. percentil; y el tercero, Q_3, es el 75o. percentil. Un software computarizado de estadística por lo general está programado para identificar cuartiles y otros cuantiles, por ejemplo los deciles, que dividen una distribución en 10 grupos de igual tamaño.

La tabla 2-9 presenta la distribución de notas en un examen de mitad de curso (X) e ilustra la utilidad de los cuartiles. En este grupo de 20 estudiantes, el 25 por ciento más bajo (o las cinco notas más bajas) son $X = 69$ y menos, el siguiente cuarto de estudiantes es de $X = 72$ a 84, el tercer cuarto son de $X = 85$ a 91, y el cuarto más alto es de $X = 93$ y más. También podemos ver que un cuarto de los estudiantes obtuvo calificaciones de 69 o menos y no obtuvo una C; la mitad obtuvo calificaciones arriba de 84, tres cuartos calificaron 91 o menos, la mitad obtuvo calificaciones entre 72 y 91, y así sucesivamente.

TABLA 2-9 | Cuartiles de una distribución de calificaciones de un examen de mitad de curso

	Especificaciones		Cálculo		
	Calificación de examen (X)	f	Porcentaje f	Porcentaje acumulado (%) f	Ubicación de los cuartiles (Q)
	31	1	5.0%	5.0	
	58	1	5.0	10.0	
	63	1	5.0	15.0	
	68	1	5.0	20.0	
$Q_1 \rightarrow$	69	1	5.0	25.0	$\leftarrow Q_1$ = 25o. percentil
	72	1	5.0	30.0	
	76	1	5.0	35.0	
	77	1	5.0	40.0	
	82	1	5.0	45.0	
$Q_2 \rightarrow$	84	1	5.0	50.0	$\leftarrow Q_2$ = 50o. percentil
	85	1	5.0	55.0	
	86	2	10.0	65.0	
	88	1	5.0	70.0	
$Q_3 \rightarrow$	91	1	5.0	75.0	$\leftarrow Q_3$ = 75o. percentil
	93	2	10.0	85.0	
	94	1	5.0	90.0	
	95	1	5.0	95.0	
	97	1	5.0	100.0	
	Total	20	100.0%		

Por último, es importante recordar ordenar las calificaciones de una distribución antes de calcular los cuartiles.

> **Cuantiles** Puntuaciones que separan una porción de los casos de una distribución.
>
> **Rango percentilar** Entre los casos en una distribución de puntuaciones, es el porcentaje de casos que caen en o debajo de un valor especificado de *X*.
>
> **Cuartiles** Cuantiles que identifican los valores de puntuaciones que dividen una distribución en cuatro grupos de igual tamaño.

Agrupación de datos de intervalo/razón

A veces, para lograr mayor claridad en la presentación de una tabla o gráfica, las distribuciones de intervalo/razón se agrupan o "colapsan" en un número más pequeño de categorías ordinales. Por ejemplo, en un estudio de adultos en Estados Unidos, los valores de la variable edad variarán de 20 a alrededor de 100, lo cual da 80 puntuaciones. Es confuso presentar una tabla que presente la frecuencia y frecuencia de porcentaje de las 80 edades. Para mayor claridad, combinamos edades en categorías de 10 años, como se ve en la tabla 2-10.

TABLA 2-10 | La variable de razón de edad, agrupada en categorías ordinales de 10 años

Especificaciones			Cálculos
Código ordinal	Grupo de edad	f	Porcentaje
1	20-29	47	10.49
2	30-39	68	15.18
3	40-49	106	23.66
4	50-59	96	21.43
5	60-69	53	11.83
6	70-79	45	10.04
7	80-89	24	5.36
8	90 y mayor	9	2.01
	Total	448	100.00

Hay algo importante qué ver respecto del agrupamiento de datos de intervalo/razón. Cuando agrupamos datos, eliminamos detalles, lo cual produce un *error de agrupamiento.* Por ejemplo, la tabla 2-10 muestra que hubo 106 encuestados entre 40 y 49 años que contestaron en la encuesta. Pero ¿la mayoría estaba más cerca de 40 o de 49? No tenemos manera de saberlo con sólo observar las puntuaciones agrupadas. Si se dispone de los datos no agrupados, podemos usarlos para obtener cálculos de promedios más precisos. En general, en cualquier momento nos movemos de un nivel de medición "superior" a otro "inferior" (esto es, de razón a intervalo, de intervalo a ordinal, de ordinal a nominal), perdemos información y ello limita lo que matemáticamente puede hacerse.

No obstante, al leer el trabajo de otros, tal vez se nos presenten datos agrupados sin los estadísticos correspondientes. En tales situaciones, el estudiante debe intentar comunicarse con el autor para obtener los estadísticos descriptivos. Si esto no es posible, entonces pueden calcularse promedios y otros estadísticos a partir de los datos agrupados, pero dichos estadísticos incluirán error de agrupamiento. Véase un estudio de Freund y Simon (1991:70) acerca del cálculo de estadísticos con datos agrupados.

Insensatez y falacias estadísticas: la importancia de tener una muestra representativa

El tamaño muestral y la representatividad muestral son cosas separadas. Una muestra grande no garantiza una muestra representativa. Una equivocación repetitiva y sistemática en el muestreo puede producir una muestra grande pero sesgada. Un caso clásico de error sistemático de muestreo ocurrió en la campaña presidencial de 1936, en la que la revista *Literary Digest* seleccionó una muestra grande de números telefónicos y propietarios de automóvil. Los resultados mostraron un abrumador apoyo para el candidato republicano, Alf Landon, sobre Franklin D. Roosevelt, el candidato demócrata. Cuando llegó el día de elecciones, no fue Landon sino Roosevelt el que ganó y nada menos que con un triunfo aplastante. La encuesta *Literary Digest* sistemáticamente no hizo caso de electores sin teléfono ni automóvil y así falló en encuestar en forma adecuada a los pobres, quienes fueron el grueso del apoyo de Roosevelt (Babbie 1992: 192-93). Hay métodos para verificar la representatividad de una muestra, y los trataremos en el capítulo 10. Por ahora basta decir que una pequeña muestra representativa es mejor que una grande no representativa. Una cucharadita de condimento picante con todos los ingredientes es una mejor prueba de gusto que una cucharada sacada de sólo una parte de la olla.

RESUMEN

1. El error estadístico es el grado conocido de imprecisión en los procedimientos empleados para recolectar y procesar información. El análisis estadístico comprende controlar el error y, así, saber si una conclusión es acertada.
2. Las dos principales fuentes de error son el error de muestreo y el error de medición.
3. Una población es un grupo grande de personas de interés particular que deseamos estudiar y entender. Muestreamos la población, reunimos datos de ella y calculamos estadísticos sobre estos datos para obtener estimaciones de parámetros poblacionales y sacar conclusiones acerca de la población.
4. Las estimaciones estadísticas están basadas en métodos científicos lógicos en los que los errores de muestreo y de medición se toman en cuenta cuando los resultados se presentan. Las estimaciones apresuradas son conclusiones mal tomadas de una población.
5. Para controlar el error de muestreo, debemos concentrarnos en sus dos fuentes específicas. La primera es el tamaño muestral. Cuanto mayor es la muestra, menor es el rango de error. Dado un tamaño muestral, la teoría de probabilidad nos permite decir exactamente con qué frecuencia es que una estadística muestral predecirá correctamente un parámetro. La segunda fuente de error de muestreo es la representatividad muestral. Una muestra no representativa puede llevar a conclusiones erróneas.
6. Para controlar el error de medición, primero debemos especificar una definición operacional y determinar el nivel de medición de una variable.
7. El nivel nominal es el nivel más sencillo de medición. *Nominal* proviene de la palabra latina para referirse al *nombre*. Una variable nominal es aquella que se mide simplemente al dar nombre a categorías.
8. Una variable ordinal es aquella con categorías con nombre y la propiedad adicional de permitir que las categorías se ordenen de más alta a más baja, de la mejor a la peor, o de la primera a la última. Las variables con lugares numéricos, por ejemplo: 1o., 2o. o 3o., también son variables ordinales.
9. Las variables de intervalo tienen las características de las variables nominales y ordinales, más una unidad numérica definida o "intervalo" de medida. Con su unidad establecida de medida, por ejemplo pulgadas, libras o millas, una variable de intervalo se califica numéricamente y está sujeta a numerosos cálculos matemáticos informativos.
10. Las variables de razón tienen las características de las variables de intervalo, más un punto cero real, donde una calificación de cero significa *ninguno*. El punto cero real permite una mayor flexibilidad de cálculos que la que se logra con las variables de intervalo. Con una variable de razón, podemos calcular razones, es decir, la cantidad de una observación en relación con otra.

11. Mientras que los códigos de las variables nominales y ordinales simplemente indican una diferencia en categoría, grupo, calidad o clase, los códigos de intervalo y las variables de razón identifican diferencias en cantidad, cantidad, grado o distancia.
12. Es importante distinguir lo que es el nivel de medición (nominal, ordinal, intervalo o razón) de lo que es la unidad de medida (pulgadas, libras, etcétera).
13. Un registro o guía de codificación es una descripción concisa de los símbolos que significa cada puntuación de cada variable en un conjunto de datos. Toda variable se codifica siguiendo los principios de inclusividad y exclusividad.
14. Una forma eficiente de dominar el carácter de los datos para una variable, es decir, la forma en que las puntuaciones difieren entre sujetos de una población, es organizar los datos en una distribución de frecuencia, una distribución de frecuencia de proporción y una distribución de frecuencia de porcentaje. Los porcentajes son más fáciles de interpretar que las proporciones. Para obtener la distribución de frecuencia de porcentaje, multiplica por 100 cada una de las frecuencias proporcionales.
15. Una frecuencia de porcentaje acumulada se utiliza para identificar cuantiles, es decir, puntuaciones que separan una porción de los casos de una distribución. Los rangos percentilares y los cuartiles son cuantiles que por lo general se presentan. Es importante ordenar las puntuaciones de una variable antes de calcular cuantiles.

EXTENSIONES DEL CAPÍTULO EN EL SITIO WEB *THE STATISTICAL IMAGINATION*

Las Extensiones del capítulo 2 del material de texto disponible en el sitio web *The Statistical Imagination*, en www.mhhe.com/ritchey2 incluye: *a*) un estudio de cuándo una variable ordinal puede ser tratada como si fuera un nivel de medición de intervalo. Estas variables ordinales "semejantes a intervalos" se pueden usar con procedimientos estadísticos más avanzados para variables de intervalo/razón. *b*) Un estudio de cómo establecer la validez y confiabilidad de una medición son consideraciones importantes para reducir al mínimo el error de medición. *c*) Una ilustración de por qué es obligatorio que el error de medición y el de muestreo se minimicen. Las mediciones perfectas son inútiles con un muestreo deficiente y viceversa.

FÓRMULAS PARA EL CAPÍTULO 2

Cálculo de la frecuencia proporcional de una categoría o puntuación:

$$p\text{ [de la muestra total } (n) \text{ en una categoría]} = \frac{f \text{ de categoría}}{n} = \frac{\#\text{ en categoría}}{n}$$

Cálculo de la frecuencia porcentual de una categoría o puntuación:

$$\%\text{ [de una muestra total } (n) \text{ en una categoría]}$$
$$= (p\text{ [de una muestra total } (n) \text{ en una categoría]})\ (100)$$

Cálculo de percentiles y cuartiles:

Elaborar una hoja de cálculo con los siguientes encabezados: