6. INFERENCIA PARA DISTRIBUCIONES

WILLIAM S. GOSSET

¿Qué podría explicar que el jefe de producción de la famosa fábrica de cerveza Guinness de Dublín, Irlanda, no sólo utilizara la estadística sino que además inventara nuevos métodos estadísticos? El anhelo de mejorar la calidad de la cerveza, por supuesto.

William S. Gosset (1876-1937) empezó a trabajar en 1899 como técnico en la fábrica de cerveza Guinness, justo después de licenciarse en la Oxford University. Muy pronto empezó a realizar experimentos y se dio cuenta de la necesidad de utilizar la estadística para comprender los resultados de éstos. ¿Cuáles son las mejores variedades de cebada y de lúpulo para producir cerveza? ¿Cómo se tienen que cultivar? ¿Cómo se deben secar y almacenar? Los resultados de los experimentos de campo, como puedes adivinar, variaban. La inferencia estadística permite descubrir la pauta que esta variación deja oculta. A principios de siglo, los métodos de inferencia se reducían a una versión de las pruebas z para las medias —incluso los intervalos de confianza eran desconocidos.

En su trabajo, Gosset se enfrentó con el problema que hemos señalado al utilizar el estadístico z: no conocía la desviación típica poblacional σ. Es más, en los experimentos de campo se obtenían pocas observaciones, por lo que la simple substitución de σ por s en el estadístico z y la suposición de que el resultado era aproximadamente normal, no daba unas conclusiones suficientemente precisas. En consecuencia, Gosset se planteó la pregunta clave, ¿cuál es la distribución exacta del estadístico $(\bar{x} - \mu)/s$?

En 1907, Gosset ya era el responsable de la investigación que se desarrollaba en Guinness. Además, Gosset también había encontrado la respuesta a la pregunta anterior y había calculado una tabla de números críticos de su nueva distribución, a la que llamamos distribución t. La nueva prueba t identificó la mejor variedad de cebada y Guinness, rápidamente, adquirió toda la semilla disponible. Guinness permitió que publicara sus descubrimientos, pero no con su propio nombre. Gosset utilizó el nombre "Student", y, en su honor, la prueba t es llamada a veces "t de Student". El trabajo estadístico le ayudó a llegar a ser jefe de producción, una posición más interesante que la de catedrático de estadística.

6.1 Introducción

Una vez vistos los principios de la inferencia estadística, podemos pasar a la práctica. Este capítulo describe los intervalos de confianza y las pruebas de significación para la media de una sola población y para la comparación de las medias de dos poblaciones. Una sección optativa discute una prueba aplicada a la comparación de las desviaciones típicas de dos poblaciones. En capítulos posteriores se describirán procedimientos de inferencia aplicados a las proporciones poblacionales, a la comparación de las medias de más de dos poblacionales y al estudio de la relación entre variables.

6.2 Inferencia para la media de una población

Los intervalos de confianza y las pruebas de significación para la media μ de una población normal se basan en la media muestral $\bar{x}$. La media de la distribución de $\bar{x}$ es μ (es decir, $\bar{x}$ es un estimador insesgado de la μ desconocida). La dispersión de $\bar{x}$ depende del tamaño de la muestra y de la desviación típica poblacional σ. En el capítulo 5 hicimos el supuesto, poco realista, de que conocíamos el valor de σ. En la práctica, σ es desconocida. Por tanto, tenemos que estimar σ a partir de los datos, incluso si nuestro principal interés es μ. La necesidad de estimar σ cambia algunos detalles de las pruebas de significación y de los intervalos de confianza para μ, pero no su interpretación.

He aquí los supuestos de los que partimos al hacer inferencia para la media poblacional:

SUPUESTOS DE LA INFERENCIA PARA LA MEDIA

- Nuestros datos son una **muestra aleatoria simple** de tamaño n de una población.
- Las observaciones proceden de una población que tiene una **distribución normal** con media μ y desviación típica σ. Los parámetros μ y σ son desconocidos.

En esta situación, la media muestral $\bar{x}$ tiene una distribución normal con media μ y desviación típica $\sigma/\sqrt{n}$. Debido a que no conocemos σ, la estimaremos a partir de la desviación típica muestral s. A continuación estimaremos la desviación típica de $\bar{x}$ a partir de $s/\sqrt{n}$. Este valor se llama *error típico* de la media muestral $\bar{x}$.

ERROR TÍPICO

Cuando la desviación típica de un estadístico se estima a partir de los datos, el resultado se llama **error típico** del estadístico. El error típico de la media muestral $\bar{x}$ es $s/\sqrt{n}$.

6.2.1 Distribuciones *t*

Cuando conocemos el valor de σ, basamos los intervalos de confianza y las pruebas para μ en la media muestral estandarizada

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Este estadístico z tiene una distribución normal estandarizada $N(0, 1)$. Cuando no conocemos σ, sustituimos la desviación típica de $\bar{x}$, $\sigma/\sqrt{n}$, por su error típico $s/\sqrt{n}$. El estadístico que se obtiene *no* tiene una distribución normal. Su distribución, que es nueva para nosotros, se llama *distribución t*.

ESTADÍSTICO *t* DE UNA SOLA MUESTRA Y DISTRIBUCIONES *t*

Obtén una muestra aleatoria simple de tamaño n de una población que tenga una distribución normal con media μ y desviación típica σ. El **estadístico *t* de una sola muestra**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

tiene una **distribución *t*** con $n - 1$ grados de libertad.

Grados de libertad

El estadístico t tiene la misma interpretación que cualquier estadístico estandarizado: indica a qué distancia se encuentra $\bar{x}$ de la media μ, expresada en desviaciones típicas. Existe una distribución t distinta para cada tamaño de muestra. Concretamos una distribución t determinada, dando sus ***grados de libertad.*** Los grados de libertad del estadístico t de una sola muestra se obtienen a partir de la desviación típica muestral s en el denominador de t. En el capítulo 1 vimos que s tenía $n - 1$ grados de libertad. Existen otros estadísticos t con diferentes grados de libertad, algunos de los cuales describiremos más adelante en este capítulo. Indicaremos, de forma abreviada, una distribución t con k grados de libertad como $t(k)$.

La figura 6.1 compara la curva de densidad de la distribución normal estandarizada con las curvas de densidad de dos distribuciones t con 2 y 9 grados de libertad, respectivamente. La figura ilustra las siguientes características de las distribuciones t:

- La forma de las curvas de densidad de las distribuciones t es similar a la forma de la curva normal estandarizada. Todas ellas son simétricas, con centro en cero, y tienen forma de campana.
- La dispersión de las distribuciones t es algo mayor que la dispersión de la distribución normal estandarizada. Las distribuciones t de la figura 6.1 tienen más probabilidad en las colas y menos en el centro que la normal estandarizada. Esto es debido a que la sustitución del parámetro fijo σ por el estadístico s introduce más variación en el estadístico t.

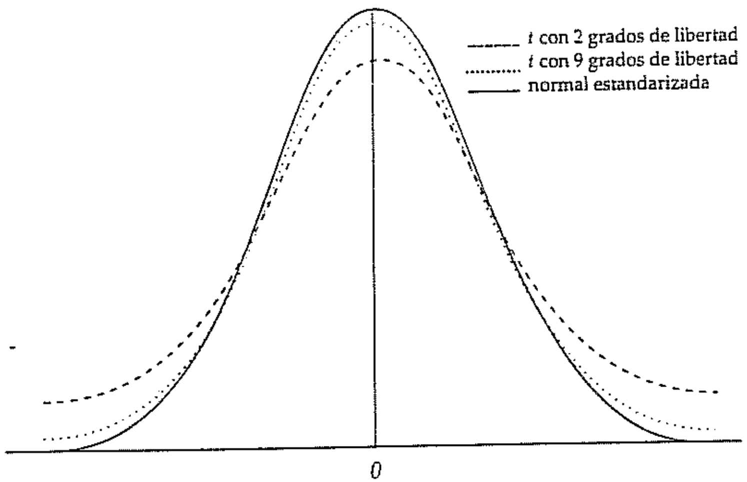


Figura 6.1. Curvas de densidad de dos distribuciones t con 2 y 9 grados de libertad, respectivamente, y de la distribución normal estandarizada. Todas son simétricas con centro en *0*. Las distribuciones t tienen más probabilidad en las colas que la distribución normal estandarizada.

- A medida que aumentan los grados de libertad k de t, la curva de densidad de $t(k)$ se parece más a la curva de densidad de una normal estandarizada $N(0, 1)$. Esto es así porque, a medida que aumenta el tamaño de la muestra, la estimación de σ a partir de s se va haciendo más precisa. Por tanto, la utilización de s en lugar de σ causa poca variación adicional cuando la muestra es grande.

La tabla C, que se encuentra en la parte final de este libro, da algunos valores críticos para algunas distribuciones t. Cada fila de la tabla contiene los valores críticos de una de las distribuciones t; los grados de libertad aparecen a la izquierda de cada fila. Para mayor comodidad, las columnas se denominan según las probabilidades p de los valores críticos superiores, es decir, la probabilidad de la cola de la derecha necesaria para las pruebas de significación, y según los niveles de confianza C (en porcentaje) necesarios para los intervalos de confianza. Ya has utilizado los valores críticos de la distribución normal estandarizada que están situados en la última fila de la tabla. Puedes comprobar, examinando cualquier columna, cómo a medida que aumentan los grados de libertad de t, los valores críticos de t se aproximan cada vez más a los valores críticos de una distribución normal estandarizada. Como en el caso de la tabla normal, los programas estadísticos a menudo hacen innecesaria la utilización de la tabla C.

EJERCICIOS

6.1. Los resultados de cuatro amigos en la prueba LSA (*Law School Aptitude Test*) tienen una media $\bar{x} = 589$ y una desviación típica $s = 37$. ¿Cuál es el error típico de la media muestral?

6.2. ¿Qué valor crítico t^* de la tabla C cumple cada una de las siguientes condiciones?:

(a) La distribución t con 5 grados de libertad tiene una probabilidad 0,05 a la derecha de t^*.

(b) La distribución t con 21 grados de libertad tiene una probabilidad 0,99 a la izquierda de t^*.

6.3. ¿Qué valor crítico t^* de la tabla C cumple cada una de las siguientes condiciones?:

(a) El estadístico t de una sola muestra de 15 observaciones tiene una probabilidad 0,025 a la derecha de t^*.

(b) El estadístico t de una sola muestra aleatoria simple de una muestra de 20 observaciones tiene una probabilidad 0,75 a la izquierda de t^*.

6.2.2 Intervalos y pruebas t

Para analizar muestras de poblaciones normales con σ desconocida, basta con sustituir la desviación típica de $\bar{x}$, $\sigma/\sqrt{n}$, por su error típico, $s/\sqrt{n}$, en los procedimientos z descritos en el capítulo 5. Los procedimientos z se convierten, entonces, en los *procedimientos t de una sola muestra*. Utiliza los valores P o los valores críticos de la distribución t con $n-1$ grados de libertad en lugar de los valores normales. La justificación y los cálculos de los procedimientos t de una sola muestra son similares a los

PROCEDIMIENTOS t DE UNA SOLA MUESTRA

Obtén una muestra aleatoria simple de tamaño n de una población de media μ desconocida. Un intervalo de confianza de nivel C para μ es

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

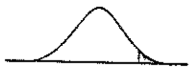
donde t^* es el valor crítico superior $(1-C)/2$ de la distribución $t(n-1)$. Este intervalo es exacto cuando la distribución de la población es normal y es aproximadamente correcto para muestras grandes en los demás casos.

Para contrastar la hipótesis $H_0: \mu = \mu_0$ a partir de una muestra aleatoria simple de tamaño n, calcula el estadístico t de una sola muestra

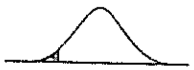
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

En términos de la variable T que tiene una distribución $t(n-1)$, el valor P para contrastar H_0 en contra de

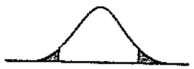
$H_a: \mu > \mu_0$ es $P = P(T \geq t)$



$H_a: \mu < \mu_0$ es $P = P(T \leq t)$



$H_a: \mu \neq \mu_0$ es $P = 2P(T \geq |t|)$



Estos valores P son exactos si la distribución de la población es normal y son aproximadamente correctos para muestras grandes en los demás casos.

procedimientos z del capítulo 5. Por tanto, nos centraremos en la utilización práctica de los procedimientos t.

EJEMPLO 6.1

Para estudiar el metabolismo de los insectos, unos investigadores alimentaron unas cucarachas con soluciones azucaradas. Después de 2, 5 y 10 horas, los investigadores diseccionaron algunas de las cucarachas y analizaron el contenido de azúcar en varios de sus tejidos.[1] Después de 10 horas, los contenidos de D-glucosa (en microgramos) en los intestinos de cinco cucarachas que se alimentaron con una solución que contenía D-glucosa, eran los siguientes:

55.95 68.24 52.73 21.50 23.78

Los investigadores calcularon un intervalo de confianza del 95% para el contenido medio de D-glucosa en los intestinos de las cucarachas en las condiciones anteriores.

Primero calcularon que

$$\bar{x} = 44.44 \quad \text{y} \quad s = 20.741$$

Los grados de libertad son $n-1=4$. En la tabla C encontramos que para un intervalo de confianza del 95%, $t^* = 2.776$. El intervalo de confianza es:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 44.44 \pm 2{,}776 \frac{20.741}{\sqrt{5}} =$$
$$= 44.44 \pm 25{,}75 =$$
$$= (18.69,\ 70{,}19)$$

La comparación de esta estimación con las estimaciones para otros tejidos y para diferentes momentos de disección permitió saber más sobre el metabolismo de las cucarachas y sobre nuevos métodos para eliminar las cucarachas de las casas y los restaurantes. El hecho de que el error de estimación sea grande se debe a que la muestra es pequeña y a que la dispersión es relativamente grande, lo que se refleja en la magnitud de s. ■

El intervalo de confianza t de una sola muestra tiene la forma

$$\text{estimación} \pm t^* \mathrm{ET}_{\text{de la estimación}}$$

[1] D. L. Shankland, *et al.*, 1968, "The effect of 5-thio-D-glucose on insect development and its absorption by insects", *Journal of Insect Physiology*, 14, págs. 63-72.

donde "ET" significa "error típico". Encontraremos diversos intervalos de confianza que tienen esta misma forma. Al igual que con los intervalos de confianza, las pruebas t son muy parecidas a las pruebas z que vimos anteriormente. He aquí un ejemplo. En el capítulo 5 utilizamos la prueba z para estos datos. Para ello, tuvimos que suponer, de forma poco realista, que conocíamos la desviación típica σ de la población. Ahora podemos hacer un análisis más realista.

EJEMPLO 6.2

Los fabricantes de refrescos prueban nuevas fórmulas para evitar la pérdida de dulzura de los refrescos *light* durante el almacenamiento. Unos catadores experimentados evalúan la dulzura de los refrescos antes y después del almacenamiento. He aquí las pérdidas de dulzura (dulzura antes del almacenamiento menos dulzura después del almacenamiento) halladas por 10 catadores para una nueva fórmula de refresco.

2.0 0.4 0.7 2.0 −0.4 2.2 −1.3 1.2 1.1 2.3

Estos datos, ¿constituyen una buena evidencia de que el refresco perdió dulzura durante el almacenamiento?

Paso 1: Hipótesis. Existen diferencias sobre la percepción de la pérdida de dulzura por parte de los catadores. Por este motivo, planteamos las hipótesis como la pérdida media de dulzura μ de una gran población de catadores. La hipótesis nula establece que la "pérdida es nula" y la hipótesis alternativa que "existe una pérdida".

$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$

Paso 2: Estadístico de contraste. Los estadísticos básicos son:

$$\bar{x} = 1{,}02 \quad \text{y} \quad s = 1{,}196$$

El estadístico t de una sola muestra es:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1{,}02 - 0}{1{,}196/\sqrt{10}} = 2{,}70$$

Paso 3: Valor P. El valor P para $t = 2{,}70$ es el área situada a la derecha de 2,70 por debajo de la curva de la distribución t con $n - 1 = 9$ grados de libertad (gl). La figura 6.2 muestra esta área. Sin la ayuda de un ordenador no podemos hallar el valor exacto de P. De todas formas, utilizando la tabla C podemos situar P entre dos valores. Busca en la fila correspondiente a 9 grados de libertad entre qué valores se

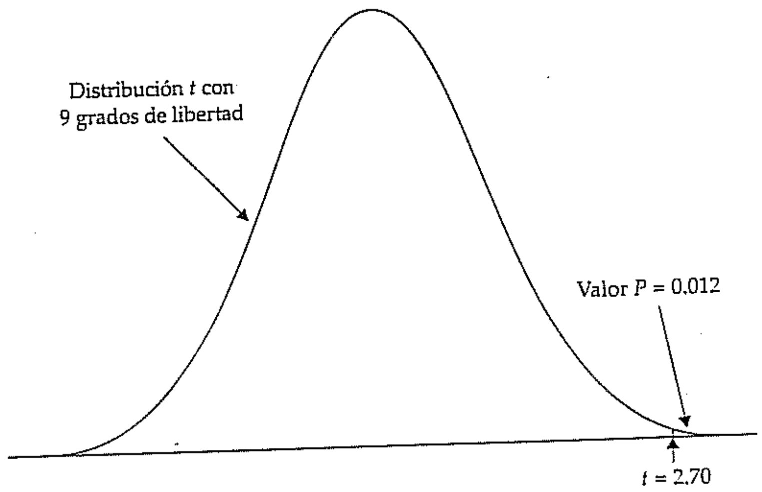


Figura 6.2. El valor P de la prueba t de una cola del ejemplo 6.2.

encuentra $t = 2.70$. Debido a que el valor observado de t se encuentra entre los valores críticos de 0.02 y 0.01, el valor P se halla entre 0.01 y 0.02. Un programa estadístico da el valor más exacto, $P = 0{,}012$. Existe una evidencia bastante fuerte a favor de que se produce una pérdida de dulzura. ■

El intervalo de confianza t del ejemplo 6.1 y la prueba t del ejemplo 6.2 se basan en supuestos razonables, pero no todos estos supuestos son fáciles de comprobar. Los dos ejemplos son experimentos. En ambos casos los investigadores tomaron especiales precauciones para evitar sesgos. Las cucarachas fueron asignadas al azar a las distintas soluciones azucaradas y a los distintos momentos de disección. Desde cualquier otro punto de vista, todas las cucarachas fueron tratadas exactamente igual. Los catadores trabajaron en cabinas aisladas para evitar que se influyeran mutuamente. En consecuencia, podemos fiarnos de los resultados obtenidos en estos dos experimentos concretos.

El análisis estadístico se basa en dos supuestos: el muestreo aleatorio y la distribución normal de las poblaciones. Tenemos que tratar a las cucarachas y a los catadores como muestras aleatorias simples de poblaciones grandes si queremos sacar conclusiones, en general, sobre las cucarachas o sobre los catadores. Las cucarachas fueron escogidas al azar de una población de cucarachas criadas en laboratorio con fines experimentales. Todos los catadores tenían una formación similar. Aunque, en realidad, no tengamos muestras aleatorias simples de las poblaciones en las que estamos interesados, estamos dispuestos a actuar como si lo fueran. Ésta es una cuestión a juzgar en cada caso.

El supuesto de que la distribución de la población es normal no se puede comprobar eficazmente con sólo 5 o 10 observaciones. En parte, los investigadores confían en la experiencia con variables similares. También examinan los datos. La figura 6.3 representa los diagramas de tallos de las dos distribuciones (los datos de las cucarachas se han redondeado). La distribución de las diferencias de dulzura de los 10 catadores no tiene una forma regular, pero no se observan espacios vacíos ni observaciones atípicas ni otros signos de falta de normalidad. Los datos de las cucarachas, por otro lado, presentan un gran espacio vacío entre las dos observaciones menores y las tres mayores. Con datos observacionales, esto podría indicar la existencia de dos tipos de cucarachas. En este caso sabemos que todas las cucarachas proceden de una misma población que se ha criado en laboratorio. El espacio vacío se debe a la variación del azar en muestras muy pequeñas.

Debido a que los procedimientos t son tan comunes, todos los programas estadísticos pueden hacerte los cálculos. La figura 6.4 muestra los resultados obtenidos a partir de tres programas estadísticos: el Data Desk, el Minitab y el S-PLUS. En cada caso, hemos introducido en el ordenador los 10 datos sobre la pérdida de dulzura como valores de una variable llamada "Refresco" y pedimos al programa que haga la prueba t de una sola muestra para H_0: $\mu = 0$, en contra de H_a: $\mu > 0$. Los resultados de los tres programas estadísticos proporcionan una información ligeramente distinta, aunque todos incluyen los cálculos básicos: $\bar{x} = 1.02$, $t = 2.70$, $P = 0.012$. Estos resultados son los mismos que hallamos en el ejemplo 6.2.

```
(a)                      (b)
      2 | 2 4                 -1 | 3
      3 |                     -0 | 4
      4 |                      0 | 4 7
      5 | 3 6                  1 | 1 2
      6 | 8                    2 | 0 0 2 3
      7 |                      3 |
      8 |
```

Figura 6.3. Diagramas de tallos de los datos del ejemplo 6.1 en (a) y del ejemplo 6.2 en (b).

EJERCICIOS

6.4. ¿Qué valor crítico t^* de la tabla C utilizarías en un intervalo de confianza para la media poblacional en cada una de las siguientes situaciones?

Data Desk

```
refresco:
Test Ho: mu(refresco) = 0 vs Ha: mu(refresco) > 0
Sample Mean = 1.02000 t-Statistic = 2.697 w/9 df
Reject Ho at Alpha = 0.0500
p = 0.0123
```

Minitab

```
TEST OF MU = 0.000 VS MU G.T. 0.000

              N      MEAN     STDEV   SE MEAN        T    P VALUE
refresco     10     1.020     1.196     0.378     2.70      0.012
```

S-PLUS

```
data:   refresco
t = 2.6967, df = 9, p-value = 0.0123
alternative hypothesis: true mean is greater than 0
sample estimates:
  mean of x
       1.02
```

Figura 6.4. Resultados de la prueba t de una sola muestra del ejemplo 6.2 de tres programas estadísticos distintos. Puedes localizar fácilmente los cálculos básicos en los resultados de los tres programas.

(a) Un intervalo de confianza del 95% basado en $n = 10$ observaciones.

(b) Un intervalo de confianza del 99% de una muestra aleatoria simple con 20 observaciones.

(c) Un intervalo de confianza del 80% de una muestra de tamaño 7.

6.5. El estadístico t de una sola muestra para contrastar

$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$

de una muestra de $n = 15$ observaciones tiene el valor $t = 1.82$

(a) ¿Cuántos grados de libertad tiene este estadístico?

(b) Da los dos valores críticos t^* de la tabla C entre los que se encuentra t. ¿Cuáles son las probabilidades p de la cola de la derecha de estos dos valores?

(c) ¿Entre qué dos valores se encuentra el valor P de la prueba?

(d) ¿Es significativo el valor $t = 1.82$ a un nivel del 5%? ¿Es significativo a un nivel del 1%?

6.6. El estadístico t de una sola muestra de $n = 25$ observaciones para contrastar la prueba de dos colas de

$$H_0 : \mu = 64$$
$$H_a : \mu \neq 64$$

tiene un valor $t = 1{,}12$.

(a) ¿Cuántos grados de libertad tiene t?

(b) Sitúa los dos valores críticos t^* de la tabla C entre los que se encuentra t. ¿Cuáles son las probabilidades p de la cola de la derecha de estos dos valores?

(c) ¿Entre qué dos valores se encuentra el valor P de la prueba? (Ten en cuenta que H_a tiene dos colas).

(d) El valor $t = 1{,}12$, ¿es estadísticamente significativo a un nivel del 10%? ¿Y a un nivel del 5%?

6.7. El envenenamiento con el pesticida DDT causa temblores y convulsiones. En un estudio sobre envenenamiento con DDT, unos investigadores suministraron una determinada cantidad de DDT a un grupo de ratas. Más tarde se tomaron datos sobre sus sistemas nerviosos para averiguar cómo causa el envenenamiento con DDT esos temblores. Una variable importante era el "periodo absolutamente refractario", es decir, el tiempo que necesita un nervio para recuperarse después de un estímulo. Este periodo, normalmente, varía. Las mediciones hechas en cuatro ratas dieron los siguientes datos (en milisegundos).[2]

1,6 1,7 1,8 1,9

(a) Halla la media del periodo absolutamente refractario $\bar{x}$ y su error típico.

(b) Da un intervalo de confianza del 90% de la media del periodo absolutamente refractario para todas las ratas de este tipo que fueron sometidas al mismo tratamiento.

6.8. El nivel de determinadas sustancias en la sangre de los enfermos del riñón sometidos a diálisis tiene que ser vigilado, ya que la insuficiencia renal y la diálisis pueden causar problemas de nutrición. Una investigadora analizó la sangre de varios pacientes sometidos a diálisis en seis visitas consecutivas. Una de las variables que se midió fue el nivel de fosfato en la sangre. El nivel de fosfato de un individuo tiende a variar normalmente a lo largo del tiempo. Los datos de uno de los pacientes, en miligramos de fosfato por decilitro de sangre, son[3]

5,6 5,1 4,6 4,8 5,7 6,4

[2] D. L. Shankland, 1964, "Involvement of spinal cord and peripheral nerves in DDT-poisoning syndrome in albino rats", *Toxicology and Applied Pharmacology*, 6, págs. 197-213.

[3] Joan M. Susic, *Dietary Phosphorus Intakes, Urinary and Peritoneal Phosphate Excretion and Clearance in Continuous Ambulatory Peritoneal Dialysis Patients*, Tesis de Licenciatura, Purdue University, 1985.

(a) Calcula la media muestral $\bar{x}$ y también su error típico.

(b) Utiliza los procedimientos t para dar un intervalo de confianza del 90% de la media del nivel de fosfato del paciente.

6.9. Supón que se sabe que la media del periodo absolutamente refractario de los nervios de las ratas no envenenadas es de 1,3 milisegundos. El envenenamiento con DDT debería retrasar la recuperación del nervio y, por tanto, aumentar este periodo. Los datos del ejercicio 6.7, ¿proporcionan suficiente evidencia a favor de esta suposición? Plantea H_0 y H_a, y lleva a cabo una prueba t. ¿Entre qué valores de la tabla C se halla el valor P? ¿Cuáles son tus conclusiones sobre esta prueba?

6.2.3 Procedimientos t para diseños experimentales por pares

Diseños por pares

El estudio del ejemplo 6.1 estimó el contenido medio de azúcar en los intestinos de unas cucarachas, aunque, para tener una visión más general, los investigadores compararon después los resultados en varios tejidos y en distintos momentos de la disección. La prueba de cata del ejemplo 6.2 corresponde a un estudio por pares, en el cual los mismos 10 catadores valoraron la dulzura de los refrescos antes y después de su almacenamiento. Los estudios comparativos son más convincentes que las investigaciones basadas en una sola muestra. Por este motivo, la inferencia basada en una sola muestra es menos habitual que la inferencia comparativa. Un diseño experimental bastante frecuente utiliza los procedimientos t de una sola muestra para comparar dos tratamientos. En un ***diseño por pares***, los sujetos se agrupan por pares y cada sujeto del par recibe uno de los dos tratamientos. El investigador puede lanzar una moneda al aire para asignar los tratamientos a los sujetos de cada par. Otra situación a la que conviene aplicar un diseño por pares es cuando hay observaciones del antes y el después con los mismos sujetos, como en la prueba de cata del ejemplo 6.2.

PROCEDIMIENTOS t EN DISEÑOS POR PARES

Para comparar las respuestas de los dos tratamientos en un diseño por pares, aplica los procedimientos t de una sola muestra a las diferencias observadas.

El parámetro μ en un procedimiento t de un diseño por pares es la media de las diferencias entre las respuestas a los dos tratamientos de los sujetos de cada par en toda la población.

EJEMPLO 6.3

Para mejorar los conocimientos de los profesores de enseñanza media de lenguas extranjeras se suelen organizar cursos de verano. En uno de estos cursos participaron veinte profesores de Francés durante cuatro semanas. Al inicio del curso, los participantes pasaron una prueba para evaluar su nivel de francés (el llamado MLA, *Modern Language Association's listening test of understanding of spoken French*). Después de las cuatro semanas, los participantes en el curso volvieron a pasar la prueba (las dos pruebas eran distintas, de manera que la realización de una prueba no tenía ninguna influencia sobre la siguiente). La tabla 6.1 muestra los resultados obtenidos por cada uno de los participantes en cada una de las pruebas. La puntuación máxima que se puede obtener en esta prueba es de 36 puntos.[4]

Tabla 6.1. Puntuación en la prueba MLA para 20 profesores de Francés.

Profesor	Prueba inicial	Prueba final	Mejora	Profesor	Prueba Inicial	Prueba final	Mejora
1	32	34	2	11	30	36	6
2	31	31	0	12	20	26	6
3	29	35	6	13	24	27	3
4	10	16	6	14	24	24	0
5	30	33	3	15	31	32	1
6	33	36	3	16	30	31	1
7	22	24	2	17	15	15	0
8	25	28	3	18	32	34	2
9	32	26	−6	19	23	26	3
10	20	26	6	20	23	26	3

Para analizar estos datos, resta la puntuación obtenida en la primera prueba de la obtenida en la segunda, y así conocer la mejora experimentada por cada profesor. Estas 20 diferencias forman una sola muestra y aparecen en la columna "mejora" de la tabla 6.1. Así, por ejemplo, el primer profesor obtuvo 32 puntos en la primera prueba y 34 puntos en la segunda, la mejora es $34 - 32 = 2$.

[4] Datos proporcionados por Joseph Wipf, Departamento de Lenguas Extranjeras y Literatura, Purdue University.

Paso 1: Hipótesis. Para valorar si los profesores mejoraron significativamente su nivel de francés, contrastamos las siguientes hipótesis:

$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$

Aquí μ es la media de las mejoras que se habrían alcanzado si todos los profesores de Francés de enseñanza media hubieran participado en el curso. La hipótesis nula establece que no hay mejora y H_a establece que la puntuación de la segunda prueba es, en promedio, mayor.

Paso 2: Estadístico de contraste. Las 20 diferencias tienen

$$\bar{x} = 2.5 \text{ y } s = 2.893$$

En consecuencia, el estadístico t de una sola muestra es

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{2.5 - 0}{2.893/\sqrt{20}} = 3.86$$

Paso 3: Valor P. Halla el valor P de la distribución $t(19)$ (recuerda que los grados de libertad son iguales al tamaño de la muestra menos 1). La tabla C muestra que 3,86 se encuentra entre los valores críticos superiores 0.001 y 0.0005 de la distribución $t(19)$. El valor P se encuentra, por tanto, entre estos dos valores. Un programa estadístico da el valor $P = 0{,}00053$. La mejora en el nivel de francés es muy poco probable que sea debida sólo al azar. Tenemos una sólida evidencia de que el curso fue efectivo para mejorar los resultados. En las publicaciones se suelen omitir los detalles de los procedimientos estadísticos rutinarios. En consecuencia, es muy posible que esta prueba se publicara de la manera siguiente: "la mejora en los resultados resultó significativa ($t = 3{,}86$, gl = 19, $P = 0{,}00053$)".

Un intervalo de confianza del 90% de la mejora media de toda la población necesita del valor crítico $t^* = 1{,}729$ de la tabla C. El intervalo de confianza es

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 2.5 \pm 1.729 \frac{2{,}893}{\sqrt{20}} =$$
$$= 2.5 \pm 1.12 =$$
$$= (1.38,\ 3.62)$$

La mejora media estimada es de 2,5 puntos, con un error de estimación de 1,12 puntos, para un nivel de confianza del 90%. Aunque estadísticamente significativo, el efecto del curso fue bastante pequeño. ■

El ejemplo 6.3 ilustra cómo expresar los datos de una prueba por pares como si fueran datos de una sola muestra. Para ello basta con calcular las diferencias entre cada par. En realidad, estamos haciendo inferencia sobre una sola población, la población de todas las diferencias entre cada par. Es incorrecto ignorar los pares y analizar los datos como si tuviéramos dos muestras, una de los profesores que participaron en el curso y otra de los profesores que no participaron. Los procedimientos inferenciales para comparar dos muestras se basan en el supuesto de que las dos muestras se obtienen de forma independiente. Este supuesto no se cumple cuando los mismos sujetos son considerados dos veces. El análisis adecuado depende del diseño utilizado para obtener los datos.

¿Qué ocurre con los supuestos de muestreo aleatorio simple y de normalidad? La utilización de los procedimientos t en el ejemplo 6.3 es un poco cuestionable. En primer lugar, estos profesores no son una muestra aleatoria simple de la población de profesores de Francés de secundaria, ya que existe un sesgo de selección a favor de los profesores más inquietos, más sensibilizados con la necesidad de mejorar su nivel de francés, que están dispuestos a sacrificar cuatro semanas de sus vacaciones de verano. Por tanto, no queda claro a qué población, exactamente, se pueden aplicar estos resultados. Esta falta de definición es habitual cuando no se ha obtenido realmente una muestra aleatoria simple de una población.

En segundo lugar, la observación de los datos muestra que algunos de los profesores obtuvieron, en la primera prueba, puntuaciones muy próximas a 36, la puntuación máxima. Estos profesores no podían mejorar mucho sus puntuaciones incluso si su nivel de francés hubiera mejorado sustancialmente. Éste es un punto débil de la prueba que se ha utilizado como instrumento de medida en el experimento. Las diferencias en los resultados podrían no indicar de manera adecuada la efectividad del curso. Ésta es una de las razones que explica por qué el aumento medio era pequeño.

Una última dificultad a la que deben enfrentarse los procedimientos t del ejemplo 6.3 consiste en que los datos muestren desviaciones de la normalidad. En un análisis por pares, la población de las *diferencias* tiene que tener una distribución normal ya que los procedimientos t se aplican a las diferencias. La figura 6.5 muestra un diagrama de tallos con las 20 diferencias. Uno de los profesores llegó a perder 6 puntos entre la primera y la segunda prueba. Sólo este sujeto bajó la media muestral de 2,95 para los restantes 19 sujetos a 2,5 para los 20 sujetos. El diagrama de tallos muestra esta observación atípica, así como la existencia de un vacío entre 3 y 6 que podría ser debido al azar. El hecho de que todas las hojas del diagrama de tallos sean ceros nos recuerda que sólo son posibles valores enteros en los resultados. Aunque eliminemos la observación atípica, la distribución sigue siendo no normal. Esta falta de normalidad, ¿nos impide utilizar la prueba t? El comportamiento de los procedimientos t cuando la población no tiene una distribución normal es una de sus características más importantes. Más adelante retomaremos este tema.

−6	0
−5	
−4	
−3	
−2	
−1	
−0	
0	000
1	00
2	000
3	000000
4	
5	
6	00000

Figura 6.5. Diagrama de tallos correspondiente a la mejora del nivel de francés de los 20 profesores del ejemplo 6.3. Todas las hojas son 0, ya que todas las puntuaciones son números enteros que se han utilizado como tallos.

EJERCICIOS

A partir de ahora, muchos ejercicios te piden que des el valor P de una prueba t. Si tienes una calculadora adecuada o un ordenador con un programa estadístico adecuado, da el valor P exacto. Si no lo tienes, utiliza la tabla C para dar los dos valores entre los que se halle P.

6.10. Un experimento agrícola compara el rendimiento de dos variedades comerciales de tomates. Unos investigadores dividen por la mitad 10 parcelas situadas en distintas localidades y plantan cada variedad de tomate en cada una de las mitades de las parcelas. Después de la cosecha, los investigadores comparan los rendimientos, en kilos por planta, en cada localidad. Las 10 diferencias (Variedad A − Variedad B) dan $\bar{x} = 0{,}34$ y $s = 0{,}83$. ¿Existe suficiente evidencia de que la Variedad A tiene un rendimiento medio mayor?

(a) Describe con palabras qué es el parámetro μ en este contexto.

(b) Plantea H_0 y H_a.

(c) Halla el estadístico t y da el valor P. ¿Cuáles son tus conclusiones?

6.11. El diseño de mandos e instrumentos influye en la facilidad con que la gente puede utilizarlos. El proyecto de un alumno consistió en investigar este efecto pidiendo a 25 alumnos diestros que hicieran girar un mando giratorio (con la mano derecha) que por la acción del giro desplazaba un indicador. Había dos instrumentos idénticos, uno con el mando que giraba hacia la derecha y otro con el mando que giraba hacia la izquierda. La siguiente tabla da los tiempos en segundos que tardó cada sujeto en desplazar el indicador una determinada distancia.[5]

Sujeto	Giro hacia la derecha	Giro hacia la Izquierda	Sujeto	Giro hacia la derecha	Giro hacia la izquierda
1	113	137	14	107	87
2	105	105	15	118	166
3	130	133	16	103	146
4	101	108	17	111	123
5	138	115	18	104	135
6	118	170	19	111	112
7	87	103	20	89	93
8	116	145	21	78	76
9	75	78	22	100	116
10	96	107	23	89	78
11	122	84	24	85	101
12	103	148	25	88	123
13	116	147			

(a) Cada uno de los 25 estudiantes que participaron en el experimento utilizó ambos instrumentos. Comenta brevemente cómo utilizarías la aleatorización para preparar el experimento.

(b) El proyecto esperaba demostrar que las personas diestras utilizan más fácilmente los mandos que giran hacia la derecha. ¿Cuál es el parámetro μ de una prueba t por pares? Plantea H_0 y H_a en términos de μ.

(c) Lleva a cabo una prueba con tus hipótesis. Da el valor P e informa de tus conclusiones.

6.12. Da un intervalo de confianza del 90% para la media de la ganancia del tiempo empleado con los mandos que giran hacia la derecha en relación al tiempo empleado con los mandos que giran hacia la izquierda, en el contexto del ejercicio 6.11. ¿Crees que el tiempo ahorrado tendría una importancia práctica si la tarea se efectuara muchas veces –por ejemplo, por parte de un trabajador de una cadena de montaje? Para ayudarte a contestar a esta pregunta, halla el tiempo medio empleado con los mandos que giran hacia la derecha como porcentaje del tiempo medio empleado con los mandos que giran hacia la izquierda.

[5] Datos proporcionados por Timothy Sturm.

6.2.4 Robustez de los procedimientos t

Los procedimientos t de una sola muestra sólo son completamente exactos cuando la población es normal. Pero las poblaciones reales nunca son exactamente normales. Por tanto, la utilidad de los procedimientos t, en la práctica, depende de cómo se vean afectados por la falta de normalidad.

> PROCEDIMIENTOS ROBUSTOS
>
> Un intervalo de confianza o una prueba de significación son considerados **robustos** si el nivel de confianza o el valor P no cambian mucho cuando se violan los supuestos en los que se basa el procedimiento.

Debido a que las colas de las curvas normales descienden muy rápidamente, las muestras de poblaciones normales deben tener muy pocas observaciones atípicas. Las observaciones atípicas sugieren que los datos no constituyen una muestra de una población normal. De forma similar a lo que ocurre con $\bar{x}$ y con s, *los procedimientos t se ven muy influidos por las observaciones atípicas*. Si elimináramos la observación atípica del ejemplo 6.3, el estadístico t cambiaría de $t = 3.86$ a $t = 5.98$ y el valor P sería mucho más pequeño. En este caso, la observación atípica provoca que el resultado de la prueba sea *menos* significativo y que el error de estimación del intervalo de confianza sea *mayor* de lo que sería sin la observación atípica. Los resultados de los procedimientos t del ejemplo 6.3 son conservadores en el sentido de que las conclusiones señalan un efecto más pequeño del que ocurriría sin la presencia de la observación atípica.

Afortunadamente, los procedimientos t son bastante robustos respecto a la falta de normalidad de la población cuando no hay observaciones atípicas, especialmente cuando la distribución de los datos es aproximadamente simétrica. Las muestras grandes mejoran la precisión de los valores P y de los valores críticos de las distribuciones t cuando la población no es normal. La principal razón para este comportamiento es el teorema del límite central. El estadístico t utiliza la media muestral $\bar{x}$, la cual es más normal a medida que aumenta el tamaño de la muestra, incluso cuando la población no tiene una distribución normal.

Siempre que tengas muestras pequeñas, antes de utilizar los procedimientos t, dibuja un gráfico para detectar asimetrías o la presencia de observaciones atípicas. Si dispones de 15 o más observaciones, los procedimientos t que hemos visto se pueden aplicar de forma segura a no ser que entre los datos existan observaciones atípicas o que la distribución sea muy asimétrica. En relación al ejemplo 6.3, en el caso de que esté justificada la eliminación de la observación atípica (podría ocurrir, por ejemplo, que el profesor que obtuvo una puntuación más baja en la segunda prueba estuviera enfermo cuando pasó la prueba), se puede utilizar el procedimiento t con las restantes 19 observaciones. He aquí unas reglas prácticas para hacer inferencia para una sola media.[6]

UTILIZACIÓN DE LOS PROCEDIMIENTOS t

- Excepto en el caso de muestras pequeñas, el supuesto de que los datos sean una muestra aleatoria simple de la población de interés es más importante que el supuesto de que la distribución de la población sea normal.
- *Tamaño de muestra menor que 15.* Utiliza los procedimientos t si los datos son aproximadamente normales. Si los datos son claramente no normales, o si existen observaciones atípicas, no utilices los procedimientos t.
- *Tamaño de muestra mayor o igual a 15.* Los procedimientos t se pueden utilizar a no ser que existan observaciones atípicas o que la distribución sea muy asimétrica.
- *Muestras grandes.* Los procedimientos t se pueden utilizar incluso para distribuciones muy asimétricas cuando la muestra sea grande, aproximadamente cuando $n \geq 40$.

EJEMPLO 6.4

Considera algunos de los conjuntos de datos que representamos gráficamente en el capítulo 1. La figura 6.6 muestra los histogramas.

- La figura 6.6(a) es un histograma de los porcentajes de residentes mayores de 65 años en cada uno de los Estados de EE UU. *Tenemos datos de toda la población de los 50 Estados; por tanto, no tiene sentido hacer inferencia.* Podemos calcular de manera exacta la media de la población. No tenemos la incertidumbre que se presenta cuando sólo disponemos de una muestra de la población, por lo que no es necesario ningún intervalo de confianza ni ninguna prueba de significación.

[6] Estas recomendaciones se basan en un extenso trabajo con ordenador. Consulta, por ejemplo, Harry O. Posten, 1979, "The robustness of the one-sample t-test over the Pearson system", *Journal of Statistical Computation and Simulation*, 9, págs. 133-149, y E. S. Pearson y N. W. Please, 1975, "Relation between the shape of population distribution and the robustness of four simple test statistics", *Biometrika*, 62, págs. 223-241.

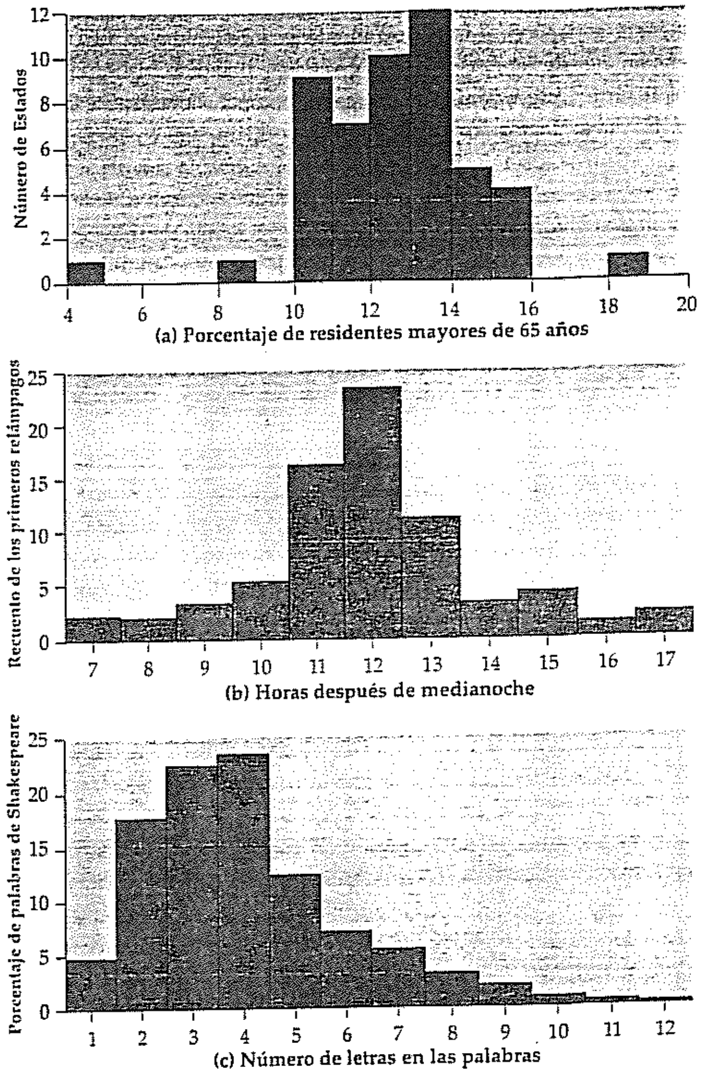


Figura 6.6. ¿Podemos utilizar los procedimientos t con estos datos? (a) Porcentaje de residentes mayores de 65 años en los distintos Estados de EE UU. *No*: se trata de toda la población, no de una muestra. (b) Hora del día en la que se produce el primer relámpago en una localidad de Colorado. *Sí*: hay más de 70 observaciones con una distribución simétrica. (c) Longitud de las palabras en las obras de Shakespeare. *Sí, si la muestra es lo suficientemente* grande como para contrarrestar la asimetría hacia la derecha.

- La figura 6.6(b) muestra la distribución de la hora del día en que se produce el primer relámpago en una región montañosa de Colorado, EE UU. Los datos contienen más de 70 observaciones que tienen una distribución simétrica. Puedes utilizar los procedimientos t para sacar conclusiones sobre la media de la hora del día en la que se produce el primer relámpago con toda fiabilidad.

- La figura 6.6(c) muestra que la distribución de la longitud de las palabras de las obras de Shakespeare es muy asimétrica hacia la derecha. No sabemos el número de observaciones de que disponemos. Se pueden utilizar los procedimientos t con una distribución de este tipo si el número de observaciones es igual o mayor que 40.

EJERCICIOS

6.13. La prueba ARSMA (*Acculturation Rating Scale for Mexican Americans*) mide el grado de adopción de la cultura anglosajona por parte de los estadounidenses de origen mexicano. Durante la etapa de elaboración del ARSMA, se sometió a la prueba a un grupo de 17 mexicanos. Sus puntuaciones, en un intervalo posible de 1,00 a 5,00, mostraron una distribución simétrica con $\bar{x} = 1{,}67$ y $s = 0{,}25$. Debido a que resultados bajos indicarían una fuerte presencia de la cultura mexicana, estos resultados ayudaron a validar la prueba.[7]

(a) Da un intervalo de confianza del 95% para la media de los resultados de los mexicanos en la prueba ARSMA.

(b) ¿Qué supuestos exige tu intervalo de confianza? ¿Cuál de estos supuestos es el más importante en este caso?

6.14. Un banco se pregunta si la eliminación de la cuota anual de la tarjeta de crédito de los clientes que carguen en la tarjeta un mínimo de 240.000 pesetas al año haría aumentar los cargos en sus tarjetas. El banco hace esta oferta a una muestra aleatoria simple de 200 clientes con tarjeta de crédito. Posteriormente, el banco compara lo que han cargado estos clientes este año con la cantidad que cargaron el año anterior. El aumento medio es de 33.200 Pta., y la desviación típica es de 10.800 Pta.

(a) ¿Existe evidencia significativa a un nivel del 1% de que la cantidad media cargada aumenta con la oferta de eliminación de la cuota? Plantea H_0 y H_a, y lleva a cabo una prueba t.

(b) Da un intervalo de confianza del 99% para la media del aumento.

(c) La distribución de las cantidades cargadas en las tarjetas de crédito es asimétrica hacia la derecha, pero no existen observaciones atípicas debido al límite de crédito que el banco impone a cada tarjeta. La utilización de los procedimientos t está justificada en este caso, a pesar de que la distribución de la población no es normal. Explica por qué.

(d) Un observador crítico puntualiza que los clientes posiblemente habrían cargado en sus tarjetas más este año que el año pasado incluso sin la oferta del banco, debido a que la situación económica de este año es mejor que la del año pasado y los tipos de interés son más bajos. Describe brevemente el diseño de un experimento para estudiar el efecto de la eliminación de la cuota que evitaría esta crítica.

[7] I. Cuellar, L. C. Harris y R. Jasso, 1980, "An acculturation scale for Mexican American normal and clinical populations", *Hispanic Journal of Behavioral Sciences*, 2, págs. 199-217.

6.15. He aquí las mediciones (en milímetros) de una dimensión crítica de 16 cigüeñales de un motor para automóviles:

224,120	224,001	224,017	223,982	223,989	223,961
223,960	224,089	223,987	223,976	223,902	223,980
224,098	224,057	223,913	223,999		

Se supone que la dimensión crítica de los cigüeñales es de 224 mm y que la variabilidad del proceso de fabricación es desconocida. ¿Existe suficiente evidencia de que la dimensión media no es de 224 mm?

(a) Comprueba gráficamente que no haya observaciones atípicas o una fuerte asimetría que pudieran amenazar la validez de los procedimientos t. ¿A qué conclusión llegas?

(b) Plantea H_0 y H_a, y lleva a cabo una prueba t. Da el valor P (de la tabla C o de un programa estadístico). ¿A qué conclusión llegas?

6.16. Algunos propietarios de viviendas compran aparatos para detectar la presencia de radón en sus casas. ¿Qué precisión tienen estos detectores? Para contestar a esta pregunta, unos investigadores colocaron 12 detectores de radón en una cámara que contenía 105 picocuries de radón por litro. Las lecturas de los detectores fueron las siguientes.[8]

91,9	97,8	111,4	122,3	105,4	95,0
103,8	99,6	96,6	119,3	104,8	101,7

(a) Dibuja un diagrama de tallos con estos datos. La distribución es algo asimétrica hacia la derecha, pero no lo suficiente para impedir el uso de los procedimientos t.

(b) ¿Existe evidencia suficiente de que la media de las lecturas de los detectores de este tipo difiere del valor real 105? Lleva a cabo una prueba de forma detallada; luego describe de forma breve tus conclusiones.

[8] Datos proporcionados por Diana Schellenberg, Facultad de Ciencias de la Salud, Purdue University.

6.2.5 Potencia de las pruebas t*

La potencia de una prueba estadística determina su capacidad para detectar desviaciones de la hipótesis nula. En la práctica, las pruebas se realizan con la intención de mostrar que la hipótesis nula es falsa, por lo que es importante que la potencia sea alta. La potencia de la prueba t de una sola muestra en contra de un determinado valor alternativo de la media poblacional μ es la probabilidad de que la prueba rechace la hipótesis nula cuando el valor alternativo de la media sea cierto. Para calcular la potencia, partimos de un determinado nivel de significación α, habitualmente $\alpha = 0.05$.

El cálculo de la potencia exacta de una prueba t requiere que σ se estime a partir de s y es algo complejo. De todas formas, un cálculo aproximado como si σ fuera conocida suele ser adecuado para planificar un estudio. Este cálculo es muy similar al cálculo de la potencia de las pruebas z que vimos en el apartado 5.5.3.

EJEMPLO 6.5

Estamos en el invierno anterior al verano en el que tuvo lugar el curso de francés del ejemplo 6.3. El director del curso está pensando en el informe que tendrá que redactar sobre el resultado del curso. Cree que con 20 participantes será capaz de detectar una mejora media de 2 puntos en el nivel de francés de los participantes. ¿Es esto razonable?

Queremos calcular la potencia de la prueba t para

$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$

en contra de la alternativa de que $\mu = 2$ cuando $n = 20$. Para calcular la potencia, tenemos que tener una idea del valor de σ. Antes de efectuar un estudio a gran escala, se suele realizar en primer lugar un estudio piloto para éste y otros propósitos. En este caso, los cursos de francés de años anteriores tuvieron una desviación típica muestral de aproximadamente 3 puntos. Por tanto, en nuestro cálculo aproximado, supondremos que $\sigma = 3$ y $s = 3$.

Paso 1. *Escribe la regla para rechazar H_0 en términos de $\bar{x}$.* La prueba t con 20 observaciones rechaza H_0 a un nivel de significación del 5% si el estadístico t

$$t = \frac{\bar{x} - 0}{s/\sqrt{20}}$$

es mayor que el valor crítico superior del 5% para $t(19)$, que es 1,729. Tomando $s = 3$, la prueba rechaza H_0 cuando

$$t = \frac{\bar{x}}{3/\sqrt{20}} \geq 1{,}729$$

$$\bar{x} \geq 1{,}729\ 3/\sqrt{20}$$

$$\bar{x} \geq 1{,}160$$

Paso 2. *La potencia es la probabilidad de rechazar H_0 suponiendo que la hipótesis alternativa sea cierta.* Queremos la probabilidad de que $\bar{x} \geq 1{,}160$ cuando $\mu = 2$. Estandariza $\bar{x}$, tomando $\sigma = 3$, para hallar esta probabilidad:

$$P(\bar{x} \geq 1{,}160) = P\left(\frac{\bar{x} - 2}{3/\sqrt{20}} \geq \frac{1{,}160 - 2}{3\sqrt{20}}\right) =$$
$$= P(Z \geq -1{,}252) = 1 - 0{,}1056 = 0{,}8944$$

Una diferencia de 2 puntos en las medias de las puntuaciones, a un nivel de significación del 5%, se detectará en el 89% de todas las muestras posibles. El director del curso puede estar razonablemente seguro de detectar una diferencia de esta magnitud. ■

* Esta sección más avanzada no es necesaria para leer el resto del libro. Esta sección se basa en la sección 5.5.

EJERCICIOS

6.17. El banco del ejercicio 6.14 contrastó una nueva idea con una muestra de 200 clientes. El banco quiere estar suficientemente seguro de detectar un aumento medio en los cargos en las tarjetas de $\mu = 10.000$ pesetas, a un nivel de significación $\alpha = 0{,}01$. Quizás bastase una muestra de sólo $n = 50$ clientes. Halla la potencia aproximada de la prueba con $n = 50$ en contra de la alternativa $\mu = 10.000$ pesetas de la siguiente manera:

(a) ¿Cuál es el valor crítico t^* para la prueba de una cola con $\alpha = 0.01$ y $n = 50$?

(b) Escribe la regla para rechazar H_0: $\mu = 0$ en términos del estadístico t. Luego toma $s = 10.800$ (una estimación basada en los datos del ejercicio 6.14) y plantea el procedimiento de rechazo en términos de $\bar{x}$.

(c) Supón que $\mu = 10.000$ (la alternativa dada) y que $\sigma = 10.800$ (una estimación con los datos del ejercicio 6.14). La potencia aproximada es la probabilidad del suceso que hallaste en (b), calculada bajo estos supuestos. Halla la potencia. ¿Recomendarías que el banco hiciera una prueba con 50 clientes, o deberían incluirse más clientes en la prueba?

6.18. Los investigadores que hicieron el experimento del ejercicio 6.10 creían que el alto valor P se debió a una baja potencia. A dichos investigadores les gustaría poder detectar una diferencia media en los rendimientos de 0,5 kilos por planta a un nivel de significación del 0,05. Basándote en el estudio previo, utiliza 0,83 como una estimación de la σ de la población y del valor de s en las futuras muestras.

(a) ¿Cuál es la potencia de la prueba del ejercicio 6.10 con $n = 10$ en contra de la alternativa $\mu = 0{,}5$?

(b) Si el tamaño de la muestra se aumenta hasta $n = 25$ parcelas, ¿cuál será la potencia de la prueba en contra de la misma alternativa?

RESUMEN

Las pruebas y los intervalos de confianza para la media μ de una población normal se basan en la media muestral $\bar{x}$ de una muestra aleatoria simple. El teorema del límite central garantiza que estos procedimientos son aproximadamente correctos con otro tipo de distribuciones poblacionales cuando las muestras son grandes.

La media muestral estandarizada es el **estadístico z de una sola muestra**,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Cuando conocemos σ, utilizamos el estadístico z y la distribución normal estandarizada.

En la práctica, no conocemos σ. Sustituye la desviación típica $\sigma/\sqrt{n}$ de $\bar{x}$ por el **error típico** $s/\sqrt{n}$ para obtener el **estadístico t de una sola muestra**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

El estadístico t tiene una **distribución t** con $n - 1$ grados de libertad.

Existe una distribución t distinta para cada valor positivo k de **grados de libertad**. Todas las distribuciones t son simétricas con forma similar a la distribución normal estandarizada. La distribución $t(k)$ se aproxima a la distribución $N(0, 1)$ a medida que k aumenta.

Un **intervalo de confianza** de nivel C exacto para la media μ de una población normal es

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

donde t^* es el valor crítico superior $(1 - C)/2$ de una distribución $t(n - 1)$.

Las **pruebas de significación** $H_0: \mu = \mu_0$, se basan en el estadístico t. Utiliza los valores P o los niveles de significación predeterminados de la distribución $t(n - 1)$ para contrastar la hipótesis.

Utiliza estos procedimientos t de una sola muestra para analizar los datos de los **diseños por pares**. Primero tienes que calcular la diferencia dentro de cada par para obtener una sola muestra.

Los procedimientos t son relativamente **robustos** cuando la población es no normal, especialmente para tamaños de muestra grandes. Los procedimientos t son útiles con datos no normales cuando $n \geq 15$, a no ser que los datos contengan observaciones atípicas o muestren una fuerte asimetría.

EJERCICIOS DE LA SECCIÓN 6.2

Cuando un ejercicio pida un valor P, da el valor P de forma exacta si tienes una calculadora adecuada o un programa estadístico. Si no es así, utiliza la tabla C para dar dos valores entre los que se encuentra P.

6.19. El estadístico t de una sola muestra para contrastar

$$H_0 : \mu = 10$$
$$H_a : \mu < 10$$

basado en $n = 10$ observaciones tiene un valor $t = -2{,}25$.

(a) ¿Cuántos grados de libertad tiene este estadístico?

(b) ¿Entre qué dos probabilidades p de la tabla C se encuentra el valor P de la prueba?

6.20. Un fabricante de pequeños electrodomésticos contrata una empresa de investigación de mercados para estimar las ventas de sus productos al por menor. Dicha empresa obtiene la información a partir de una muestra de tiendas minoristas. Este mes, una muestra aleatoria simple de 75 tiendas pone de manifiesto que este tipo de tiendas vendieron un promedio de 24 batidoras de este fabricante, con una desviación típica de 11 batidoras.

(a) Da un intervalo de confianza del 95% para la media de las batidoras vendidas en todas las tiendas.

(b) La distribución de las ventas es muy asimétrica hacia la derecha, ya que hay muchas tiendas pequeñas y unas pocas muy grandes. La utilización de t en (a) es razonablemente segura a pesar de esta violación del supuesto de normalidad. ¿Por qué?

6.21. En un experimento comparativo aleatorizado sobre el efecto que el calcio en la dieta tiene sobre la presión sanguínea, unos investigadores separaron al azar a 54 hombres sanos en dos grupos. Un grupo recibió calcio; el otro, un placebo. Al comienzo del estudio, los investigadores midieron una serie de variables en los sujetos. El informe del estudio da $\bar{x} = 114.9$ y $s = 9.3$ para la presión sanguínea sistólica de los 27 sujetos del grupo placebo.

(a) Da un intervalo de confianza del 95% para la media de la presión sanguínea de la población de la que proceden estos sujetos.

(b) ¿Qué supuestos en relación a la población y al diseño del estudio exige el procedimiento que utilizaste en (a)? ¿Cuáles de estos supuestos son, en este caso, importantes para la validez del procedimiento?

6.22. La cromatografía de gases es una técnica utilizada para analizar pequeñas concentraciones de determinados compuestos. Los cromatógrafos de gases se calibran haciendo lecturas repetidas de un patrón con una concentración conocida. Un estudio sobre la calibración de estos aparatos utiliza un patrón que contiene 1 nanogramo por litro (es decir, 10^{-9} gramos) de un determinado compuesto. Las lecturas que se obtienen son las siguientes.[9]

21.6 20.0 25.0 21.9

Se sabe por experiencia que las lecturas varían de acuerdo con una distribución normal, a no ser que una observación atípica indique un error en el análisis. Estima la media de las lecturas obtenidas y da el error de estimación para un nivel de confianza de tu elección. Luego explica a un químico que no sepa estadística lo que significa tu error de estimación.

6.23. Los embriones de artemia pueden entrar en una fase de latencia en la cual la actividad metabólica desciende a un nivel bajo. Unos investigadores que estudiaron esta fase de latencia determinaron los niveles de varias sustancias importantes para el metabolismo. Los investigadores presentaron sus resultados en forma de tabla, con la nota, "los valores son medias ± ETM de tres muestras independientes". El valor del nivel de ATP en la tabla era 0,84 ± 0,01. Los biólogos que lean el artículo tienen que ser capaces de descifrar qué significa esto.[10]

(a) ¿Qué significa la abreviatura ETM?

(b) Los investigadores realizaron tres determinaciones del nivel de ATP, que dieron $\bar{x} = 0{,}84$. ¿Cuál fue la desviación típica muestral s de estas determinaciones?

(c) Da un intervalo de confianza del 90% para la media del nivel de ATP en los embriones de artemia en latencia.

[9] Datos del apéndice de D. A. Kurtz (ed.), *Trace Residue Analysis*, series de los Simposios de la Sociedad Americana de Química, número 284, 1985.

[10] S. C. Hand y E. Gnaiger, 1988, "Anaerobic dormancy quantified in *Artemia* embryos", *Science*, 239, págs. 1.425-1.427.

6.24. La siguiente tabla da los resultados en la prueba MLA (*Modern Language Association listening test*) de 20 profesores de Español de secundaria que participaron en un curso intensivo de Lengua Española en verano. El contexto es idéntico al descrito en el ejemplo 6.3.[11]

Sujeto	Resultado antes del curso	Resultado después del curso	Sujeto	Resultado antes del curso	Resultado después del curso
1	30	29	11	30	32
2	28	30	12	29	28
3	31	32	13	31	34
4	26	30	14	29	32
5	20	16	15	34	32
6	30	25	16	20	27
7	34	31	17	26	28
8	15	18	18	25	29
9	28	33	19	31	32
10	20	25	20	29	32

(a) Esperamos poder demostrar que la asistencia al curso mejora la comprensión del español hablado. Plantea las H_0 y H_a apropiadas. Asegúrate de que identificas el parámetro que aparece en las hipótesis.

(b) Comprueba gráficamente si existen observaciones atípicas o asimetrías claras en los datos que utilizarás en las pruebas estadísticas. Da tus conclusiones sobre la validez de la prueba.

(c) Lleva a cabo la prueba. ¿Puedes rechazar H_0 a un nivel de significación del 5%? ¿Y a un nivel de significación del 1%?

(d) Da un intervalo de confianza del 90% para la media de los incrementos de los resultados de la prueba MLA debidos a la asistencia al curso.

6.25. La prueba ARSMA (ejercicio 6.13) se comparó con una prueba similar, la prueba BI (*Bicultural Inventory*), haciendo pasar ambas pruebas a 22 ciudadanos de EE UU de origen mexicano. Las dos pruebas tienen el mismo intervalo de resultados (de 1,00 a 5,00) y se ajustaron de manera que las medias de las dos pruebas al aplicarse a los grupos experimentales que se utilizaron para desarrollarlas fueran similares. Se observó una alta correlación entre los resultados de las dos pruebas, lo que es una evidencia a favor de que las dos pruebas miden las mismas características. Los investigadores querían saber si las medias de los resultados poblacionales eran iguales en

[11] Datos proporcionados por Joseph Wipf, Departamento de Lenguas Extranjeras y Literatura, Purdue University.

las dos pruebas. Las diferencias de los resultados (ARSMA – BI) de los 22 sujetos tuvo $\bar{x} = 0{,}2519$ y $s = 0{,}2767$.

(a) Describe brevemente cómo organizarías la aplicación de las dos pruebas a los 22 sujetos. Incluye la aleatorización.

(b) Lleva a cabo una prueba de significación para la hipótesis de que las dos pruebas tienen la misma media poblacional. Da el valor P y justifica tus conclusiones.

(c) Da un intervalo de confianza del 95% para la diferencia entre los resultados medios de las dos pruebas.

6.26. Un estudio sobre el salario de los altos ejecutivos examinó los ingresos, después de tener en cuenta la inflación, de los altos ejecutivos de 104 empresas durante el periodo 1977-1988. Entre los datos había los promedios de los aumentos salariales anuales de cada uno de los 104 ejecutivos. La media de los aumentos porcentuales de los salarios era del 6,9%. Los datos mostraron una gran variación, con una desviación típica del 17,4%. La distribución era claramente asimétrica hacia la derecha.[12]

(a) A pesar de la asimetría de la distribución, no había observaciones atípicas extremas. Explica por qué podemos utilizar los procedimientos t con estos datos.

(b) ¿Cuántos grados de libertad hay? Cuando los grados de libertad exactos no estén en la tabla C, utiliza los grados de libertad con el valor menor inmediato de la tabla.

(c) Da un intervalo de confianza del 99% para el aumento medio de los salarios de los altos ejecutivos. ¿Qué condición esencial deben cumplir los datos para que los resultados sean fiables?

6.27. La tabla 1.3 ofrece las edades de los presidentes de EE UU cuando tomaron posesión del cargo. No tiene sentido utilizar los procedimientos t (o cualquier otro procedimiento estadístico) para dar un intervalo de confianza del 95% para la media de la edad de los presidentes. Explica por qué.

6.28 (Optativo). El ejercicio 6.25 habla de un pequeño estudio que compara la prueba ARSMA y la prueba BI, dos pruebas que determinan la orientación cultural de los ciudadanos de EE UU de origen mexicano. Este estudio, ¿detectaría habitualmente una diferencia en los resultados medios igual a 0,2? Para contestar a esta pregunta, calcula la potencia aproximada del contraste (con $n = 22$ sujetos y $\alpha = 0{,}05$) de

$$H_0 : \mu = 0$$
$$H_a : \mu \neq 0$$

[12] Charles W.L. Hill y Phillip Phan, 1991, "CEO tenure as a determinant of CEO pay", *The Academy of Management Journal*, 34, págs. 707-717.

en contra de la alternativa $\mu = 0{,}2$. Fíjate en que es una prueba de dos colas.

(a) ¿Cuál es el valor crítico de la tabla C para $\alpha = 0{,}05$?

(b) Escribe el procedimiento para rechazar H_0 a un nivel $\alpha = 0{,}05$. Luego, toma $s = 0{,}3$, el valor aproximado observado en el ejercicio 6.25, y expresa el criterio de rechazo en términos de $\bar{x}$.

(c) Halla la probabilidad de este suceso cuando $\mu = 0{,}2$ (la alternativa) y $\sigma = 0{,}3$ (estimada con los datos del ejercicio 6.25) mediante un cálculo de probabilidad normal. Esta probabilidad es la potencia aproximada.

6.3 Comparación de dos medias

La comparación de dos poblaciones o de dos tratamientos es una de las situaciones más comunes que hay que afrontar en estadística aplicada. A estas situaciones las llamaremos *problemas de dos muestras*.

PROBLEMAS DE DOS MUESTRAS

- El objetivo de la inferencia es la comparación de las respuestas a dos tratamientos o la comparación de las características de dos poblaciones.
- Tenemos una muestra distinta de cada población o de cada tratamiento.

6.3.1 Problemas de dos muestras

Un problema de dos muestras puede surgir de un experimento comparativo aleatorizado que divida aleatoriamente a los sujetos en dos grupos y exponga a cada grupo a un tratamiento distinto. La comparación de dos muestras aleatorias seleccionadas independientemente de dos poblaciones también es un problema de dos muestras. A diferencia de los diseños por pares que hemos estudiado anteriormente, las unidades experimentales no se agrupan por pares en las dos muestras y las dos muestras pueden ser de distinto tamaño. Los procedimientos inferenciales para los datos de dos muestras son distintos de los procedimientos inferenciales de los datos por pares. He aquí algunos problemas de dos muestras típicos.

EJEMPLO 6.6

(a) Un investigador médico está interesado en el efecto de un incremento de calcio en la dieta sobre la presión sanguínea. Por este motivo, el investigador lleva a cabo un experimento comparativo aleatorizado en el cual un grupo de sujetos recibe un suplemento de calcio y un grupo de control recibe un placebo.

(b) Un psicólogo desarrolla una prueba que determina la capacidad de un individuo para captar la personalidad de la gente que le rodea. El psicólogo quiere compararar esta capacidad en los estudiantes universitarios de sexo masculino con la de las de sexo femenino y, con este fin, hace pasar la prueba a un grupo numeroso de estudiantes de cada sexo.

(c) Un banco quiere saber cuál de los dos planes para incentivar la utilización de sus tarjetas de crédito es mejor. El banco aplica cada plan a una muestra aleatoria simple de sus clientes y después compara la cantidad cargada durante seis meses en las tarjetas de crédito de los sujetos de los dos grupos. ■

EJERCICIOS

6.29. Las siguientes situaciones exigen hacer inferencia sobre una o dos medias. Identifica cada situación como (1) de una sola muestra, (2) de un diseño por pares o (3) de dos muestras. Los procedimientos de la sección 6.2 se aplican a los casos (1) y (2). Estamos a punto de iniciar el estudio de los procedimientos que se aplican al caso (3).

(a) Un pedagogo quiere saber si es más efectivo plantear preguntas antes o después de introducir un nuevo concepto en un texto de matemáticas destinado a la enseñanza primaria. El pedagogo prepara dos extractos del texto que introducen el concepto; uno con preguntas para suscitar el interés de los niños antes de introducir el concepto y el otro con preguntas de repaso después de introducir el concepto. El pedagogo utiliza cada extracto del texto con un grupo distinto de niños y compara los resultados de los dos grupos de niños mediante un examen sobre este tema.

(b) Una pedagoga enfoca el mismo problema de forma diferente. Esta pedagoga preparara extractos de un texto sobre dos temas sin relación entre sí. Cada extracto tiene dos versiones, una con preguntas antes de introducir los conceptos y otra con preguntas después de introducir los conceptos. Los sujetos son un solo grupo de niños. Cada niño estudia los dos temas, uno (escogido al azar) con preguntas antes y otro con preguntas después de los nuevos conceptos. La investigadora compara los resultados de los exámenes de cada alumno en los dos temas para ver qué tema aprendió mejor.

6.30. Las siguientes situaciones exigen hacer inferencia para una o dos medias. Identifica cada situación como (1) de una sola muestra, (2) de un diseño por pares o (3) de dos muestras. Los procedimientos de la sección 6.2 se aplican a los casos (1) y (2). Estamos a punto de iniciar el estudio de los procedimientos que se aplican al caso (3).

(a) Para comprobar la fiabilidad de un nuevo método de análisis, un químico tiene un patrón con una concentración conocida de una determinada sustancia. El químico lleva a cabo 20 análisis de la concentración de la sustancia en este patrón con el nuevo método y comprueba si existe un sesgo comparando la media de los resultados con la concentración conocida.

(b) Una química comprueba la fiabilidad del mismo método de análisis utilizando otro procedimiento. Esta química no tiene ninguna muestra de referencia, pero dispone de un método analítico conocido. La química quiere saber si los resultados de los dos métodos concuerdan. Con este objetivo, la química prepara un patrón de concentración desconocida y analiza su concentración 10 veces con el nuevo método y 10 veces con el método conocido.

6.3.2 Comparación de las medias de dos poblaciones

Podemos examinar los datos de dos muestras gráficamente comparando sus diagramas de tallos (para muestras pequeñas) o bien sus histogramas o diagramas de caja (para muestras grandes). Ahora aplicaremos las ideas de la inferencia formal a esta situación. Cuando las dos distribuciones poblacionales son simétricas y, especialmente, cuando al menos son aproximadamente normales, la comparación de las respuestas medias en las dos poblaciones es el objetivo más común de la inferencia. He aquí los supuestos que haremos.

> **SUPUESTOS PARA LA COMPARACIÓN DE DOS MEDIAS**
>
> - Tenemos dos **muestras aleatorias simples** de dos poblaciones distintas. Las muestras son **independientes**. Es decir, una muestra no tiene ninguna influencia sobre la otra. Así, por ejemplo, la agrupación por pares viola la independencia. Medimos la misma variable en las dos muestras.
> - Las dos poblaciones tienen **distribuciones normales**. Las medias y las desviaciones típicas de las dos poblaciones son desconocidas.

Llama a la variable que medimos x_1 en la primera población y x_2 en la segunda, ya que la variable puede tener distribuciones distintas en las dos poblaciones. La notación que utilizaremos para describir las dos poblaciones es:

Población	Variable	Media	Desviación típica
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

Existe un total de cuatro parámetros desconocidos, las dos medias y las dos desviaciones típicas. Los subíndices nos recuerdan qué población describe cada parámetro. Queremos comparar las dos medias poblacionales, dando un intervalo de confianza de su diferencia $\mu_1 - \mu_2$ o contrastando la hipótesis de que no existen diferencias, $H_0: \mu_1 = \mu_2$.

Utilizamos las medias y las desviaciones típicas muestrales para estimar los parámetros desconocidos. De nuevo, los subíndices nos recuerdan de qué muestra procede cada estadístico. La notación que describen las muestras es:

Población	Tamaño de la muestra	Media muestral	Desviación típica muestral
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

Para hacer inferencia sobre la diferencia $\mu_1 - \mu_2$ entre las dos medias poblacionales, partimos de la diferencia entre las dos medias muestrales $\bar{x}_1 - \bar{x}_2$.

EJEMPLO 6.7

Un aumento de calcio en la dieta, ¿reduce la presión sanguínea? El examen de una gran muestra de personas reveló que existía una relación entre la ingestión de calcio en la dieta y la presión sanguínea. La relación era más fuerte en el caso de los hombres negros. A partir de los estudios observacionales no es posible deducir relaciones de causa-efecto. Por tanto, unos investigadores diseñaron un experimento comparativo aleatorizado.

Los sujetos del experimento eran 21 hombres negros sanos. Diez de estos hombres, escogidos al azar, tomaron un suplemento de calcio durante 12 semanas. Los 11 restantes, el grupo de control, tomaron un placebo de aspecto idéntico a las píldoras que tomó el otro grupo. El experimento fue doblemente ciego. La variable respuesta es la disminución en la presión sistólica de la sangre de los sujetos después de 12 semanas, en milímetros de mercurio. Una respuesta negativa indica un aumento de la presión sanguínea.[13]

[13] Roseann M. Lyle, *et al.*, 1987, "Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men", *Journal of the American Medical Association*, 257, págs. 1.772-1.776.

El Grupo 1 es el que recibió el calcio y el Grupo 2 el que recibió el placebo. He aquí los datos de los 10 hombres del Grupo 1,

$$7 \quad -4 \quad 18 \quad 17 \quad -3 \quad -5 \quad 1 \quad 10 \quad 11 \quad -2$$

y los datos de los 11 hombres del Grupo 2,

$$-1 \quad 12 \quad -1 \quad -3 \quad 3 \quad -5 \quad 5 \quad 2 \quad -11 \quad -1 \quad -3$$

A partir de los datos, calcula los estadísticos resumen:

Grupo	Tratamiento	n	$\bar{x}$	s
1	Calcio	10	5.000	8.743
2	Placebo	11	− 0.273	5.901

El Grupo 1 muestra una disminución de la presión sanguínea, $\bar{x}_1 = 5.000$, mientras que el grupo del placebo no experimentó prácticamente ningún cambio, $\bar{x}_2 = -0.273$. Estos resultados, ¿proporcionan evidencia suficiente de que el calcio disminuye más la presión sanguínea de la población de hombres negros sanos que el placebo? ■

El ejemplo 6.7 corresponde a una comparación de dos muestras. Escribimos las hipótesis en términos de la disminución media de la presión sanguínea que observaríamos en toda la población, μ_1 para los hombres que toman calcio durante 12 semanas y μ_2 para los hombres que toman el placebo. Las hipótesis son

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

Queremos contrastar estas hipótesis y también estimar la ventaja del calcio sobre el placebo, $\mu_1 - \mu_2$.

¿Se satisfacen los supuestos? Debido a la aleatorización, podemos considerar los dos grupos experimentales como dos muestras aleatorias simples independientes. Aunque las muestras son pequeñas, comprobamos que no existen signos de no normalidad importantes examinando los datos. He aquí un diagrama de tallos doble de las respuestas. (Hemos dividido los tallos. Fíjate en que las respuestas negativas obligan a que − 0 y 0 vayan en tallos distintos; fíjate también en que la ordenación de las hojas en cada tallo tiene en cuenta, por ejemplo, que − 3 es menor que − 1).

Las respuestas del grupo del placebo aparecen como aproximadamente normales. El grupo que tomó calcio tiene una distribución irregular, que es frecuente cuan-

Calcio		Placebo
	−1	1
5	−0	5
234	−0	33111
1	0	23
7	0	5
10	1	2
87	1	

do tenemos sólo pocas observaciones. Pero no hay observaciones atípicas ni desviaciones de la normalidad que impidan la utilización de los procedimientos t.

El estimador natural de la diferencia $\mu_1 - \mu_2$ es la diferencia entre las medias muestrales:

$$\bar{x}_1 - \bar{x}_2 = 5{,}000 - (-0{,}273) = 5{,}273$$

Este estadístico mide la ventaja promedio del calcio sobre el placebo. Para utilizarlo en inferencia, debemos conocer su distribución.

6.3.3 Distribución muestral de $\bar{x}_1 - \bar{x}_2$

He aquí las principales características de la distribución de la diferencia $\bar{x}_1 - \bar{x}_2$ entre las medias muestrales de dos muestras aleatorias simples independientes. Estas características se pueden deducir de forma matemática utilizando los teoremas de probabilidad, o se pueden poner de manifiesto mediante simulaciones.

- La media de $\bar{x}_1 - \bar{x}_2$ es $\mu_1 - \mu_2$. Es decir, la diferencia de las medias muestrales es un estimador insesgado de la diferencia de las medias poblacionales.
- La varianza de la diferencia es la *suma* de las varianzas de $\bar{x}_1$ y de $\bar{x}_2$, que es

$$\frac{\sigma_1^2}{n_1} \div \frac{\sigma_2^2}{n_2}$$

Fíjate en que las *varianzas* se suman, pero las desviaciones típicas no.

- Si las dos distribuciones poblacionales son normales, entonces la distribución de $\bar{x}_1 - \bar{x}_2$ también es normal.

Estadístico z de dos muestras

Debido a que el estadístico $\bar{x}_1 - \bar{x}_2$ tiene una distribución normal, lo podemos estandarizar para obtener un estadístico normal estandarizado z. Para obtener el ***estadístico z de dos muestras***, réstale su media y divide el resultado por su desviación típica:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

6.3.4 Procedimientos t de dos muestras

Estadístico t de dos muestras

No conocemos las desviaciones típicas poblacionales σ_1 y σ_2. De forma parecida a lo que hicimos en el caso de una sola muestra, sustituye, en el estadístico z de dos muestras, las desviaciones típicas $\sigma_i/\sqrt{n_i}$, por sus errores típicos $s_i/\sqrt{n_i}$. El resultado es el ***estadístico t de dos muestras***:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

La interpretación del estadístico t es la misma que la de cualquier estadístico z o t: nos indica a qué distancia se encuentra $\bar{x}_1 - \bar{x}_2$ de $\mu_1 - \mu_2$ tomando como unidad de medida la desviación típica. Desgraciadamente, el estadístico t de dos muestras *no* tiene una distribución t. Una distribución t sustituye a una distribución $N(0, 1)$ cuando reemplazamos en un estadístico z una sola desviación típica por su error típico. En este caso, sustituimos las dos desviaciones típicas por sus errores típicos correspondientes. Esto no genera un estadístico con una distribución t.

Sin embargo, en los problemas de inferencia sobre dos muestras, el estadístico t de dos muestras se utiliza con los valores críticos t. Hay dos formas de hacerlo.

Opción 1. Utiliza los procedimientos basados en el estadístico t con los valores críticos de una distribución t con grados de libertad calculados a partir de los datos. Los grados de libertad generalmente no son números enteros. De esta forma se obtiene una aproximación muy exacta de la distribución t.

Opción 2. Utiliza los procedimientos basados en el estadístico t con los valores críticos de una distribución t con grados de libertad iguales al menor de los valores $n_1 - 1$ y $n_2 - 1$. Estos procedimientos son siempre conservadores para dos poblaciones normales.

La mayoría de programas estadísticos utilizan, en los problemas de dos muestras, el estadístico t de dos muestras con la opción 1, a no ser que el usuario exija otro método. La utilización de esta opción sin la ayuda de un ordenador es algo complicada. Es por este motivo que presentamos en primer lugar la opción 2, que es más sencilla. Te recomendamos que utilices la opción 2 cuando hagas los cálculos sin la ayuda de un ordenador. Si utilizas un programa estadístico, el programa hará los cálculos, por defecto, con la opción 1. He aquí un resumen del procedimiento empleado en la opción 2 que incluye una indicación de por qué se trata de un procedimiento "conservador".

PROCEDIMIENTO t DE DOS MUESTRAS

Obtén una muestra aleatoria simple de tamaño n_1 de una población normal de media μ_1 desconocida y una muestra aleatoria simple independiente de tamaño n_2 de otra población normal de media μ_2 desconocida. El intervalo de confianza para $\mu_1 - \mu_2$ dado por

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

tiene un nivel de confianza de *al menos C*, independientemente de cuáles sean las desviaciones típicas poblacionales. Aquí, t^* es el valor crítico superior de $(1 - C)/2$ de la distribución $t(k)$, donde k es el menor de los valores $n_1 - 1$ y $n_2 - 1$.

Para contrastar la hipótesis $H_0: \mu_1 = \mu_2$, calcula el estadístico t de dos muestras

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

y utiliza los valores P o los valores críticos de la distribución $t(k)$. El verdadero valor P o nivel de significación predeterminado siempre será *igual o menor que* el valor calculado a partir de $t(k)$, independientemente de cuáles sean los valores que tengan las desviaciones poblacionales desconocidas.

Este procedimiento t de dos muestras siempre es seguro, puesto que da valores P *mayores* y niveles de confianza *menores* que los verdaderos valores. La diferencia entre los valores obtenidos y los verdaderos suele ser bastante pequeña, a no ser que el tamaño de las dos muestras sea pequeño y desigual. A medida que aumenta el tamaño de las muestras, los valores probabilísticos basados en la distribución t con grados de libertad iguales al menor de los valores $n_1 - 1$ y $n_2 - 1$ son cada vez más exactos.[14] Los siguientes ejemplos ilustran la utilización de la prueba t de dos muestras.

EJEMPLO 6.8

Los investigadores del ejemplo 6.7 pueden utilizar el procedimiento t de dos muestras para comparar el efecto del calcio y del placebo. El estadístico de contraste de la hipótesis nula $H_0: \mu_1 = \mu_2$ es

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} =$$

$$= \frac{5.000 - (-0.273)}{\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}} =$$

$$= \frac{5.273}{3.2878} = 1.604$$

La distribución t en la que basaremos el cálculo de las probabilidades tiene 9 grados de libertad, es decir, el menor de $n_1 - 1 = 9$ y $n_2 - 1 = 10$. Debido a que H_a es de una cola, la cola de la derecha, el valor P es el área situada a la derecha de $t = 1{,}604$ por debajo de la curva $t(9)$. La figura 6.7 ilustra este valor P. La tabla C muestra que este valor se encuentra entre 0,05 y 0,1. El experimento halló evidencia a favor de que el calcio reduce la presión sanguínea, pero la evidencia no llega a los valores tradicionales del 5% y del 1%.

Para un intervalo de confianza del 90%, la tabla C muestra que el valor crítico de $t(9)$ es $t^* = 1{,}833$. Tenemos una confianza del 90% de que la ventaja media del calcio sobre el placebo, $\mu_1 - \mu_2$, se encuentra en el intervalo

[14] Se puede encontrar información detallada sobre los procedimientos t conservadores en Paul Leaverton y John J. Birch, 1969, "Small sample power curves for the two-sample location problem", *Technometrics*, 11, págs. 299-307; en Henry Scheffé, "Practical solutions of the Behrens-Fisher problem", *Journal of the American Statistical Association*, 65, págs. 1.501-1.508; y en D. J. Best y J. C. W. Rayner, 1987, "Welch's approximate solution for the Behrens-Fisher problem", *Technometrics*, 29, págs. 205-210.

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} =$$

$$= [5{,}000 - (-0{,}273)] \pm 1{,}833\sqrt{\frac{8{,}743^2}{10} + \frac{5{,}901^2}{11}} =$$

$$= 5{,}273 \pm 6{,}026 =$$

$$= (-0{,}753\,,\ 11.299)$$

Que un intervalo de confianza del 90% contenga el 0 nos indica que no podemos rechazar H_0: $\mu_1 = \mu_2$, en contra de la hipótesis alternativa de dos colas a un nivel de significación $\alpha = 0{,}10$. ■

El tamaño de la muestra tiene una gran influencia sobre el valor P de una prueba. Un efecto que no sea significativo para un determinado nivel α en una muestra pequeña será significativo en una muestra mayor. A la vista de que las muestras del ejemplo 6.8 son más bien pequeñas, sospechamos que más datos podrían indicar un efecto significativo del calcio. En la publicación del estudio se combinaron estos resultados para negros con resultados para blancos y se ajustaron las diferencias previas a la prueba entre ambos grupos. Utilizando este análisis más detallado, los investigadores fueron capaces de dar un valor P igual a 0.008.

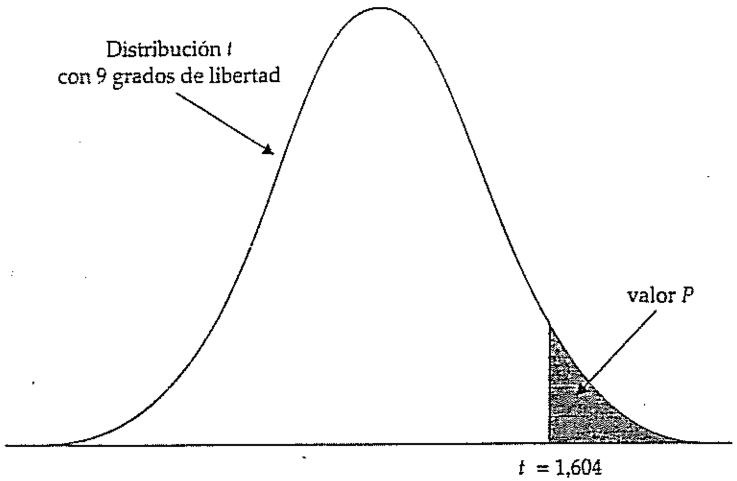


Figura 6.7. El valor P del ejemplo 6.8. Este ejemplo utiliza el método conservador, que conduce a una distribución t con 9 grados de libertad.

EJEMPLO 6.9

La prueba Chapin (*Chapin Social Insight Test*) es una prueba psicológica diseñada al objeto de determinar la precisión con la que un individuo capta la personalidad de las personas que le rodean. Los posibles resultados de la prueba van de 0 a 41. Durante el desarrollo de la prueba Chapin, ésta se aplicó a diferentes grupos de personas. He aquí los resultados obtenidos con un grupo de estudiantes universitarios de arte de ambos sexos:[15]

Grupo	Sexo	n	$\bar{x}$	s
1	Hombres	133	25.34	5.05
2	Mujeres	162	24.94	5.44

A partir de estos datos, ¿se puede afirmar que la capacidad de las mujeres y de los hombres para captar la personalidad de ls personas que les rodean es distinta?

Paso 1: Hipótesis. Debido a que antes de analizar los datos no teníamos pensado que la diferencia entre hombres y mujeres pudiese ir en una u otra dirección, escogemos una hipótesis alternativa de dos colas. Las hipótesis son

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

Paso 2: Estadístico de contraste. El estadístico t de dos muestras es

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} =$$

$$= \frac{25{,}94 - 24.34}{\sqrt{\frac{5.05^2}{133} + \frac{5.44^2}{162}}} = 0{,}654$$

Paso 3: Valor P. Hay 132 grados de libertad, el menor de

$$n_1 - 1 = 133 - 1 = 132 \ \text{ y}$$
$$n_2 - 1 = 162 - 1 = 161$$

[15] H. G. Gough, *The Chapin Social Insight Test*, Consulting Psychologists Press, Palo Alto, California, 1968.

La figura 6.8 ilustra el valor P. Hállalo comparando 0,654 con valores críticos de la distribución $t(132)$ y luego doblando p, ya que la alternativa es de dos colas. En la tabla C no aparece el valor correspondiente a 132 grados de libertad. En consecuencia, utilizaremos el valor de la tabla menor más próximo, 100 grados de libertad. La tabla C muestra que 0,654 no alcanza el valor crítico 0,25, que es la mayor probabilidad de la cola de la derecha de la tabla C. El valor P es, por tanto, mayor que 0,50. Los datos no proporcionan suficiente evidencia de que existan diferencias entre hombres y mujeres en las medias de los resultados de la prueba Chapin ($t = 0{,}654$, gl = 132, $P > 0{,}5$). ■

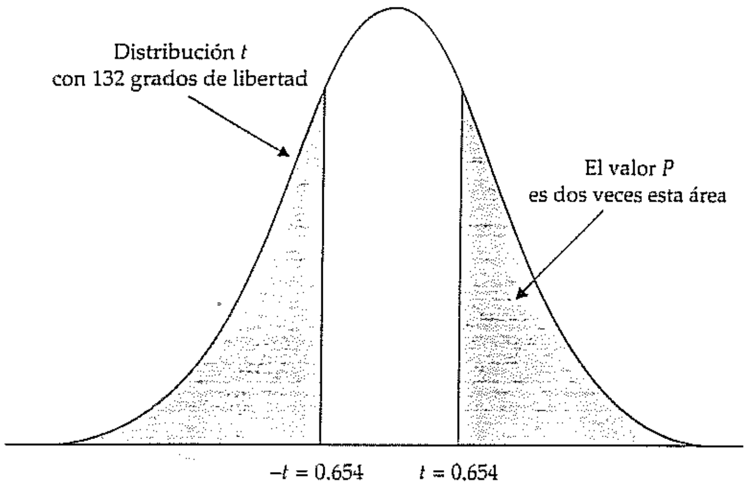


Figura 6.8. El valor P del ejemplo 6.9. Para hallar P, busca el valor del área situada a la derecha de $t = 0{,}654$ y dóblalo, ya que la alternativa es de dos colas.

El investigador del ejemplo 6.9 no llevó a cabo ningún experimento, sino que comparó muestras de dos poblaciones. Las muestras grandes permiten que el supuesto de que las poblaciones tengan distribuciones normales pierda importancia. Las medias muestrales serán, en cualquier caso, aproximadamente normales. El mayor problema que se plantea es saber a qué población se pueden aplicar las conclusiones. Estos estudiantes no son una muestra aleatoria simple de todos los estudiantes de arte del país. Si son voluntarios de una sola universidad, los resultados muestrales no se pueden extender a una población más amplia.

EJERCICIOS

6.31. En un estudio sobre cirugía del corazón se analizó el efecto de unos fármacos llamados bloqueadores beta sobre el pulso de los pacientes durante las intervenciones quirúrgicas. Los sujetos experimentales se dividieron al azar en dos grupos de 30 pacientes cada uno. A un grupo se le suministró un bloqueador beta; al otro, un placebo. El equipo quirúrgico registró el pulso de cada paciente en un momento crítico de la operación. El grupo experimental registró una media de 65.2 latidos por minuto y una desviación típica de 7,8. Para el grupo de control, la media fue de 70.3 latidos por minuto y la desviación típica de 8,3. Los datos parecen aproximadamente normales.

(a) Los bloquedores beta, ¿reducen el pulso? Plantea las hipótesis y lleva a cabo una prueba t. El resultado, ¿es significativo a un nivel del 5%? ¿Y a un nivel del 1%?

(b) Da un intervalo de confianza del 99% para la diferencia entre las medias de los pulsos de los dos tratamientos.

6.32. En un estudio sobre los daños causados por los escarabajos en los cultivos de avena, unos investigadores contaron el número de larvas de escarabajo por tallo en pequeñas parcelas plantadas con avena después de aplicar aleatoriamente uno de los dos tratamientos siguientes: ningún insecticida o malathion (un insecticida) a una dosis de 275 gramos por hectárea. Los datos aparecen como aproximadamente normales. He aquí los estadísticos de resumen.[16]

Grupo	Tratamiento	n	$\bar{x}$	s
1	Control	13	3.47	1.21
2	Malathion	14	1.36	0.52

A un nivel de significación del 1%, ¿los datos proporcionan suficiente evidencia de que el malathion reduce el número medio de larvas por tallo? Asegúrate de plantear H_0 y H_a.

6.33. Un estudio compara una muestra de empresas griegas que quebraron con una muestra de empresas griegas sin problemas. Un indicador de la salud financiera de una empresa es el cociente entre el activo y el pasivo de la empresa, A/P. El año anterior al de la quiebra, el estudio halló un A/P medio de 1.72565 en el grupo de empresas sin problemas y de 0.78640 en el grupo de empresas que quebró. El estudio afirma que $t = 7{,}36$.[17]

[16] M. C. Wilson, *et al.*, 1969, "Impact of cereal leaf beetle larvae on yields of oats", *Journal of Economic Entomology*, 62, págs. 699-702.

[17] Costas Papoulias y Panayiotis Theodossiou, 1992, "Analysis and modeling of recent business failures in Greece", *Managerial and Decision Economics*, 13, págs. 163-169.

(a) Puedes sacar conclusiones a partir de esta t sin utilizar una tabla e incluso sin conocer los tamaños de las muestras (siempre que las muestras no sean muy pequeñas). ¿Cuál es tu conclusión? ¿Por qué no necesitas conocer el tamaño de la muestra y la tabla de valores críticos de t?

(b) En realidad, el estudio investigó a 33 empresas que quebraron y a 68 empresas sin problemas. ¿Cuántos grados de libertad utilizarías para la prueba t si consideras la aproximación conservadora que se recomienda utilizar cuando no se dispone de un programa estadístico?

6.3.5 Otra vez la robustez

Los procedimientos t de dos muestras son más robustos que los procedimientos t de una sola muestra, especialmente cuando las distribuciones no son simétricas. Cuando los tamaños de las dos muestras son iguales y las dos poblaciones que se comparan tienen distribuciones similares, las probabilidades de la tabla t son bastante exactas para un amplio espectro de distribuciones, incluso cuando el tamaño de las muestras es tan pequeño como $n_1 = n_2 = 5$.[16] Cuando la forma de las distribuciones de las dos poblaciones es distinta, se necesitan muestras mayores.

Como guía práctica, adapta las indicaciones del apartado 6.2.4 para la utilización de los procedimientos t de una sola muestra a los procedimientos t de dos muestras, mediante la sustitución de "tamaño de la muestra" por "suma de los tamaños de las muestras", $n_1 + n_2$. Estas indicaciones son seguras, especialmente cuando las dos muestras son de igual tamaño. Cuando prepares un estudio de dos muestras, procura elegir, siempre que puedas, muestras de igual tamaño. Los procedimientos t de dos muestras son más robustos ante una posible falta de normalidad en este caso, y los valores de las probabilidades conservadoras son más exactos.

EJERCICIOS

6.34. En Estados Unidos muchas universidades cuentan con que los alumnos utilicen los salarios que puedan obtener durante el verano para pagarse parte del coste de la universidad. Pero, ¿son importantes estos ingresos? Una universidad estudió este tema preguntando a una muestra de estudiantes cuánto habían ganado durante el verano. Precindiendo de los estudiantes que no estuvieron empleados, hubo 1.296 respuestas. He aquí los datos de forma resumida.[19]

Grupo	n	$\bar{x}$ dólares	s dólares
Hombres	675	1.884,52	1.368,37
Mujeres	621	1.360,39	1.037,46

(a) La distribución de los ingresos es claramente asimétrica hacia la derecha. Sin embargo, la utilización de los procedimientos t está justificada. ¿Por qué?

(b) Da un intervalo de confianza del 90% para la diferencia entre la media de los ingresos de verano de los hombres y la de las mujeres.

(c) Una vez se decidió el tamaño de la muestra, ésta se escogió tomando cada vigésimo nombre de una lista alfabética de todos los estudiantes de la universidad. ¿Es razonable considerar que las muestras son muestras aleatorias simples de las poblaciones de hombres y mujeres que estudian en la universidad?

(d) ¿Qué otra información pedirías antes de aceptar que los resultados describen a todos los estudiantes?

6.35. El maíz común no tiene la cantidad del aminoácido lisina que necesitan los animales en su pienso. Unos científicos han desarrollado unas variedades de maíz que contienen una mayor cantidad de lisina. En una prueba sobre la calidad del maíz con alto contenido en lisina destinado a pienso animal, un grupo experimental de 20 pollos de un día de edad empezó a recibir una ración que contenía el nuevo maíz. Un grupo de control de otros 20 pollos recibió una ración que era idéntica a la anterior, con la excepción de que contenía maíz normal. He aquí las ganancias de peso (en gramos) de los pollos a los 21 días.[20]

Control				Experimental			
380	321	366	356	361	447	401	375
283	349	402	462	434	403	393	426
356	410	329	399	406	318	467	407
350	384	316	272	427	420	477	392
345	455	360	431	430	339	410	326

(a) Representa gráficamente los datos. ¿Hay observaciones atípicas o asimetrías claras que pudieran impedir la utilización de los procedimientos t?

[16] Consulta los extensos estudios de simulación de Harry O. Posten, 1978, "The robustness of the two-sample t-test over the Pearson system", *Journal of Statistical Computation and Simulation*, 6, págs. 295-311, y Harry O. Posten, H. Yeh y Donald B. Owen, 1982, "Robustness of the two-sample t-test under violations of the homogeneity assumption", *Communications in Statistics*, 11, págs. 109-126.

[19] Datos de 1982 proporcionados por Marvin Schlatter, División de Ayuda Financiera, Purdue University.

[20] G. L. Cromwell, *et al.*, 1968, "A comparison of the nutritive value of opaque-2, floury-2 and normal corn for the chick", *Poultry Science*, 47, págs. 840-847.

(b) ¿Existe suficiente evidencia de que los pollos alimentados con el maíz con un alto contenido en lisina ganan peso más deprisa? Lleva a cabo una prueba y da tus conclusiones.

(c) Da un intervalo de confianza del 95% para la media del peso extra de los pollos alimentados con el maíz con un alto contenido en lisina.

6.36. La encuesta SSHA (*Survey of Study Habits and Attitudes*) es una prueba psicológica que mide la motivación, la actitud hacia la universidad y los hábitos de estudio de los estudiantes. Los resultados van de 0 a 200. Una selecta universidad privada pasa la encuesta SSHA a una muestra aleatoria simple de estudiantes de primer curso de ambos sexos. Los resultados de las mujeres son los siguientes:

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

Y los de los hombres:

108	140	114	91	180	115	126	92	169	146
109	132	75	88	113	151	70	115	187	104

(a) Examina cada muestra gráficamente, prestando especial atención a las observaciones atípicas y a las asimetrías. ¿Es aceptable la utilización de un procedimiento t con estos datos?

(b) La mayoría de los estudios han hallado que la media de los resultados de la prueba SSHA de los hombres es menor que la media de los resultados de un grupo comparable de mujeres. ¿Es esto cierto para los estudiantes de primer curso de esta universidad? Lleva a cabo una prueba y da tus conclusiones.

(c) Calcula un intervalo de confianza del 90% para la diferencia entre la media de los resultados en la prueba SSHA de los hombres y la de las mujeres estudiantes de primer curso de esta universidad.

6.3.6 Procedimientos t de dos muestras más precisos*

El estadístico t de dos muestras no tiene una distribución t. Es más, la distribución exacta cambia a medida que las desviaciones típicas poblacionales desconocidas, σ_1 y σ_2, cambian. De todas formas, se dispone de una excelente aproximación.

* La lectura de esta sección se puede omitir, a no ser que quieras entender cómo calculan las probabilidades los programas estadísticos.

DISTRIBUCIÓN APROXIMADA DEL ESTADÍSTICO t DE DOS MUESTRAS

La distribución del estadístico t de dos muestras es aproximadamente una distribución t con los grados de libertad gl dados por

$$\text{gl} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Esta aproximación es bastante precisa cuando ambos tamaños muestrales n_1 y n_2 son mayores o iguales que 5.

Los procedimientos t de dos muestras son exactamente iguales a los procedimientos t que hemos visto hasta ahora, la única diferencia es que utilizamos la distribución t con gl grados de libertad para obtener los valores críticos y los valores P.

EJEMPLO 6.10

En el experimento sobre el calcio de los ejemplos 6.7 y 6.8 los datos dieron

Grupo	Tratamiento	n	$\bar{x}$	s
1	Calcio	10	5.000	8.743
2	Placebo	11	−0.273	5.901

Para ganar precisión podemos calcular los valores críticos de la distribución t con los grados de libertad dados por

$$\text{gl} = \frac{\left(\frac{8.743^2}{10} + \frac{5.901^2}{11}\right)^2}{\frac{1}{9}\left(\frac{8.743^2}{10}\right)^2 + \frac{1}{10}\left(\frac{5.901^2}{11}\right)^2} = \frac{116{,}848}{7{,}494} = 15.59$$

Fíjate en que el número de los grados de libertad gl no es un número entero.

El intervalo de confianza del 90% conservador para $\mu_1 - \mu_2$ del ejemplo 6.8, utilizó el valor crítico $t^* = 1.833$ basado en 9 grados de libertad. Un intervalo de confianza más exacto sustituye este valor crítico por el valor crítico de la distribución t con gl = 15,59 grados de libertad. No podemos hallar este valor crítico de forma exacta sin la utilización de un programa estadístico. Para hacer un cálculo aproximado, utiliza el valor menor más cercano (15 grados de libertad) de la tabla C. El valor crítico es $t^* = 1.753$. El intervalo de confianza del 90% es ahora

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} =$$

$$= [5.000 - (-0.273)] \pm 1.753\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}} =$$

$$= 5.273 \pm 5.764 = (-0.491\ ,\ 11.037)$$

Este intervalo de confianza es un poco más corto (el error de estimación es de 5,764 en vez de 6.026) que el intervalo de confianza conservador del ejemplo 6.8. ■

Tal como ilustra el ejemplo 6.10, los procedimientos t de dos muestras son exactamente iguales a los de antes, siendo la única diferencia la utilización de una distribución t con más grados de libertad.

El número gl del recuadro anterior siempre es al menos tan grande como el menor de los valores $n_1 - 1$ y $n_2 - 1$. Por otro lado, gl nunca es mayor que la suma de los dos grados de libertad individuales $n_1 + n_2 - 2$. El número de grados de libertad gl no es, generalmente, un número entero. Existe una distribución t para cualquier valor positivo de los grados de libertad, a pesar de que la tabla C contiene sólo datos correspondientes a valores enteros de los grados de libertad. Algunos programas estadísticos hallan gl y luego utilizan la distribución t que tiene el número entero positivo menor más próximo. Otros programas estadísticos tienen cuidado en utilizar t(gl) incluso si gl no es un número entero positivo. No te aconsejamos la utilización habitual de este método a no ser que el ordenador haga los cálculos. Con un ordenador, en cambio, el procedimiento más preciso no cuesta ningún esfuerzo, tal como ilustra el siguiente ejemplo.

EJEMPLO 6.11

El envenenamiento por DDT causa convulsiones en los seres humanos y en otros mamíferos. Unos investigadores quieren comprender la causa de estas convulsiones. En un experimento comparativo aleatorizado, los investigadores compararon

6 ratas envenenadas con DDT con un grupo de control de 6 ratas no envenenadas. La medida de la actividad eléctrica de los nervios es la clave para conocer la naturaleza del envenenamiento por DDT. Cuando un nervio es estimulado, se produce en él una respuesta eléctrica pronunciada seguida por una segunda respuesta menor. El experimento halló que la segunda respuesta eléctrica era mayor en las ratas envenenadas con DDT que en las ratas del grupo de control. Este hallazgo ayudó a los investigadores a comprender los mecanismos del envenenamiento por DDT.[21]

Los investigadores midieron la intensidad de la segunda respuesta al estimular un nervio de la pata de la rata, como un porcentaje de la intensidad de la primera respuesta. En las ratas envenenadas los resultados fueron

12.207 16,869 25.050 22.429 8.456 20.589

Los datos del grupo de control fueron

11.074 9,686 12.064 9.351 8.182 6.642

Las dos poblaciones son razonablemente normales, en la medida en que se puede juzgar a partir de seis observaciones. Los datos del DDT tienen una mayor dispersión que los datos del control. La diferencia entre las medias es bastante grande. De todas formas, con muestras tan pequeñas la media muestral es muy variable. Una prueba de significación puede ayudar a confirmar que estamos ante un efecto real. Debido a que los investigadores no hicieron ninguna conjetura previa sobre la intensidad de la segunda respuesta en las ratas envenenadas con DDT, utilizaremos una alternativa de dos colas:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

He aquí los resultados obtenidos con el programa estadístico SAS para estos datos.[22]

[21] El ejemplo es una adaptación libre de D. L. Shankland, 1964, "Involvement of spinal cord and peripheral nerves in DDT-poisoning syndrome in albino rats", *Toxicology and Applied Pharmacology*, 6, págs. 197-213.

[22] No utilizamos ni el Minitab ni el Data Desk en el ejemplo 6.11, ya que estos programas abrevian el procedimiento t de dos muestras. Estos programas calculan los grados de libertad gl utilizando la fórmula del recuadro del apartado 6.3.6; no obstante, luego aproximan los grados de libertad hasta el número entero menor más próximo para obtener el valor P. El resultado es ligeramente menos preciso que el valor P de la distribución t(gl).

```
                              TTEST PROCEDURE
Variable: RESPUESTA
GROUP                 N             Mean          Std Dev        Std Error
--------------------------------------------------------------------------
DDT                   6      17.60000000       6.34014839       2.58835474
CONTROL               6       9.49983333       1.95005932       0.79610839

Variances             T               DF         Prob>|T|
---------------------------------------------------------
Unequal          2.9912              5.9           0.0247
Equal            2.9912             10.0           0.0135
```

El programa SAS proporciona los resultados de dos procedimientos t: el procedimiento de dos muestras usual (suponiendo que las dos varianzas poblacionales son distintas, *"unequal variances"*) y un procedimiento especial que supone que las dos varianzas poblacionales son iguales. Estamos interesados en el primero de estos dos procedimientos. El estadístico t de dos muestras toma el valor $t = 2{,}9912$, los grados de libertad son gl = 5,9 y el valor P de la distribución $t(5{,}9)$ es 0,0247. Este resultado proporciona una clara evidencia de que la amplitud media de la segunda respuesta es mayor en las ratas envenenadas con DDT. ■

Si hubiésemos utilizado el procedimiento conservador basado en 5 grados de libertad (tanto $n_1 - 1$ como $n_2 - 1$ son 5), ¿hubiéramos obtenido un resultado distinto del obtenido en el ejemplo 6.11? El estadístico es exactamente el mismo: $t = 2.9912$. El valor P conservador es $2P(T \geq 2.9912)$, donde T tiene una distribución $t(5)$. La tabla C muestra que 2.9912 se encuentra entre los valores críticos superiores de la distribución $t(5)$, 0,02 y 0,01. En consecuencia, el valor P, en la prueba de dos colas, se encuentra entre 0.02 y 0.04. A efectos prácticos es el mismo resultado que el obtenido con el programa estadístico. Como sugiere este ejemplo y el 6.10, la diferencia entre los dos procedimientos t (el conservador y el más exacto) suele carecer de importancia práctica. Por este motivo recomendamos la utilización del procedimiento conservador, que es más sencillo, para hacer inferencia sin ordenador.

EJERCICIOS

6.37. El ejemplo 6.11 comenta un análisis de los efectos del envenenamiento con DDT. El programa estadístico utiliza la prueba t de dos muestras con los grados de libertad que da el recuadro del apartado 6.3.6. A partir de los resultados de $\bar{x}_i$ y s_i proporcionados por el ordenador, comprueba que los valores del estadístico de contraste $t = 2{,}99$ y de los grados de libertad gl = 5,9, proporcionados por el programa estadístico, sean correctos.

6.38. ¿Qué aspectos de la técnica del remo permiten distinguir entre remadores de competición principiantes y experimentados? Unos investigadores compararon dos grupos de remadores de élite: un grupo experimentado y un grupo de principiantes. Para ello analizaron diversos aspectos mecánicos del estilo de remo mientras los sujetos remaban en un ergómetro. Una variable importante es la velocidad angular de la rodilla, que describe la velocidad a la que se abre la articulación de la rodilla cuando la pierna empuja el cuerpo hacia atrás en el banco deslizante. Los datos indican que no hay ni observaciones atípicas ni fuertes asimetrías. He aquí los resultados obtenidos con el SAS.[23]

```
                              TTEST PROCEDURE
Variable: RODILLA
GROUP                 N             Mean          Std Dev        Std Error
--------------------------------------------------------------------------
EXPERIMENTADOS       10       4.18283335       0.47905935       0.15149187
PRINCIPIANTES         8       3.01000000       0.95894830       0.33903942

Variances             T               DF         Prob>|T|
---------------------------------------------------------
Unequal          3.1583              9.8           0.0104
Equal            3.3918             16.0           0.0037
```

(a) Los investigadores creían que la velocidad de la rodilla sería mayor en los deportistas experimentados. Plantea H_0 y H_a.

(b) ¿Cuál es el valor del estadístico t de dos muestras y su valor P? (Fíjate en que el SAS proporciona valores P de dos colas. Si necesitas un valor P de una cola, divide el valor de dos colas por 2). ¿Qué conclusiones obtienes?

(c) Da un intervalo de confianza del 90% para la diferencia entre las medias de las velocidades de las rodillas de los deportistas experimentados y principiantes.

6.39. Los investigadores del ejercicio anterior también querían saber si los remadores experimentados y los principiantes se diferencian en el peso o en cualquier otra característica física. He aquí los resultados obtenidos con el SAS sobre el peso de los deportistas en kilogramos.

```
                              TTEST PROCEDURE
Variable: PESO
GROUP                 N             Mean          Std Dev        Std Error
--------------------------------------------------------------------------
EXPERIMENTADOS       10      70.3700000        6.10034898       1.92909973
PRINCIPIANTES         8      68.4500000        9.03999930       3.19612240

Variances             T               DF         Prob>|T|
---------------------------------------------------------
Unequal          0.5143             11.8           0.6165
Equal            0.5376             16.0           0.5982
```

[23] Basado en W. N. Nelson y C. J. Widule, "Kinematic analysis and efficiency estimate of intercollegiate female rowers", manuscrito no publicado, 1983.

¿Existe una evidencia significativa a favor de una diferencia en las medias de los pesos de los dos grupos de deportistas? Plantea H_0 y H_a, da el estadístico t de dos muestras, su valor P y escribe tus conclusiones (fíjate en que el SAS proporciona valores P de dos colas. Si necesitas un valor P de una cola, divide el valor de dos colas por dos).

6.3.7 Procedimientos t de dos muestras con varianza común*

En el ejemplo 6.11, el programa estadístico ofreció dos posibilidades de ejecución de las pruebas t. Una llevaba el nombre en inglés de *"unequal variances"* (varianzas distintas) y otra llevaba el nombre de *"equal variances"* (varianzas iguales). El procedimiento para varianzas distintas es nuestro procedimiento t de dos muestras. Esta prueba es válida tanto si las varianzas poblacionales son iguales como si son distintas. La otra posibilidad es una versión especial del estadístico t de dos muestras que supone que las dos poblaciones tienen la misma varianza. Este procedimiento promedia las dos varianzas muestrales para estimar la varianza poblacional común. El estadístico resultante se llama el estadístico t de dos muestras con varianza común (en los programas de ordenador aparece a veces con el nombre inglés *"pooled t statistic"*). Es igual a nuestro estadístico t si el tamaño de las dos muestras es igual, pero no en caso contrario. Podríamos utilizar el estadístico t de dos muestras con varianza común en las pruebas de significación y en los intervalos de confianza.

El estadístico t de dos muestras con varianza común tiene la ventaja de que tiene exactamente una distribución t con $n_1 + n_2 - 2$ grados de libertad *si* las dos varianzas poblacionales son realmente iguales. Obviamente, las varianzas poblacionales son, a menudo, distintas. Es más, el supuesto de igualdad de varianzas es difícil de comprobar a partir de los datos. La utilización del estadístico t de dos muestras con varianza común era frecuente antes de que el uso de los ordenadores facilitara el empleo de la aproximación exacta a la distribución de nuestro estadístico t de dos muestras. Actualmente sólo es útil en situaciones especiales. No podemos utilizar la t de dos muestras con varianza común en el ejemplo 6.11, ya que está claro que la varianza del grupo de ratas envenenadas con DDT es mayor que la varianza del grupo de control.

RESUMEN

Los datos en un **problema de dos muestras** son dos muestras aleatorias simples independientes, cada una de ellas obtenida de una población distinta distribuida normalmente.

*Este apartado es un tema especial optativo.

Las pruebas de significación y los intervalos de confianza para la diferencia entre las medias μ_1 y μ_2 de dos poblaciones parten de la diferencia $\bar{x}_1 - \bar{x}_2$ entre las dos medias muestrales. Con distribuciones no normales, el teorema del límite central garantiza que los procedimientos de cálculo son aproximadamente correctos cuando las muestras son grandes.

Obtén muestras aleatorias simples independientes de tamaños n_1 y n_2 de dos poblaciones normales de parámetros μ_1, σ_1 y μ_2, σ_2, respectivamente. El **estadístico t de dos muestras** es

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

El estadístico t *no* tiene exactamente una distribución t.

Los procedimientos de inferencia conservadores para comparar μ_1 y μ_2 utilizan el estadístico t de dos muestras con la distribución $t(k)$. El valor de los grados de libertad k es el menor de $n_1 - 1$ y $n_2 - 1$. Para valores de probabilidad más exactos, utiliza la distribución t(gl) con los grados de libertad gl estimados a partir de los datos. Este procedimiento es el que se utiliza normalmente en los programas estadísticos.

El **intervalo de confianza** para $\mu_1 - \mu_2$ dado por

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

tiene un nivel de confianza de al menos C si t^* es el valor crítico superior $(1 - C)/2$ de $t(k)$, donde k es el menor de $n_1 - 1$ y $n_2 - 1$.

Las **pruebas de significación** para $H_0: \mu_1 = \mu_2$ basadas en

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

tienen un verdadero valor P que no es mayor que el calculado para $t(k)$.

Los consejos prácticos sobre el uso de los procedimientos t de dos muestras son similares a los consejos prácticos sobre el uso del estadístico t de una sola muestra. Se recomienda que el tamaño de las muestras sea igual.

EJERCICIOS DE LA SECCIÓN 6.3

En los ejercicios en los que deban utilizarse procedimientos t de dos muestras, puedes emplear como grados de libertad el más pequeño de $n_1 - 1$ y $n_2 - 1$ o el valor gl más exacto dado en el recuadro del apartado 6.3.6. Te recomendamos la primera opción, a no ser que utilices un ordenador. Muchos de estos ejercicios te piden que reflexiones sobre la aplicación práctica de la estadística, además de llevar a cabo los procedimientos t.

6.40. Unos "buscadores de talentos" sometieron a la prueba SAT (*Scholastic Assessment Test*), pensada para jóvenes que han terminado sus estudios de secundaria, a muchachos de 13 años. Entre 1980 y 1982, participaron en las pruebas 19.883 muchachos y 19.937 muchachas. Los resultados medios de los dos sexos en la prueba de lengua son casi iguales, pero hay una clara diferencia entre ambos sexos en la prueba de matemáticas. No se conoce cuál es la razón de esta diferencia. He aquí los datos.[24]

Grupo	$\bar{x}$	s
Muchachos	416	87
Muchachas	386	74

Da un intervalo de confianza del 99% de la diferencia entre la media de los resultados de los muchachos y la media de los resultados de las muchachas de la población. Los resultados de la prueba SAT, ¿tienen que tener una distribución normal para que tu intervalo de confianza sea válido? ¿Por qué?

6.41. Un estudio sobre deficiencias de hierro en bebés comparó muestras de bebés cuyas madres escogieron distintas formas de alimentarlos. Un grupo estaba formado por bebés a los que se les dio el pecho. Los bebés del otro grupo se alimentaron con una leche en polvo sin ningún suplemento de hierro. He aquí un resumen de los niveles de hemoglobina en la sangre de los sujetos experimentales a los 12 meses de edad.[25]

Grupo	n	$\bar{x}$	s
Leche materna	23	13,3	1,7
Leche en polvo	19	12,4	1,8

[24] De un anuncio en *Science*, 224 (1983), págs. 1.029-1.031.

[25] M. F. Picciano y R. H. Deering, 1980, "The influence of feeding regimens on iron status during infancy", *The American Journal of Clinical Nutrition*, 33, págs. 746-753.

(a) ¿Existe evidencia significativa de que el nivel medio de hemoglobina es diferente en los bebés alimentados con leche materna? Plantea H_0 y H_a, y lleva a cabo una prueba t. Da el valor P. ¿Qué conclusiones obtienes?

(b) Da un intervalo de confianza del 95% de la diferencia entre las medias del nivel de hemoglobina de las dos poblaciones de bebés.

(c) Describe los supuestos en los que se basan los procedimientos que has utilizado en (a) y (b).

(d) Este estudio, ¿es un experimento? ¿Por qué? ¿Cómo afecta esto a las conclusiones que extraemos del estudio?

6.42. La buena forma física está relacionada con ciertas características de la personalidad. En un estudio sobre esta relación, el profesorado de mediana edad de una universidad que había participado voluntariamente en un programa atlético fue dividido, mediante un reconocimiento médico, en dos grupos. En un grupo, los que estaban en buena forma y en el otro grupo los que estaban en baja forma. Posteriormente, los sujetos pasaron la prueba CSPFQ (*Cattell Sixteen Personality Factor Questionnaire*) para determinar su personalidad. He aquí los datos sobre la "fuerza de personalidad" de cada sujeto.[26]

Grupo	Forma	n	$\bar{x}$	s
1	Baja	14	4,64	0,69
2	Buena	14	6,43	0,43

(a) La diferencia entre las medias de "fuerza de personalidad" de los dos grupos, ¿es significativa a un nivel del 5%? ¿Y a un nivel del 1%? Asegúrate de plantear H_0 y H_a.

(b) ¿Puedes extender estos resultados a la población de todos los hombres de mediana edad? Explica por qué.

6.43. Una empresa de investigación de mercados proporciona a unos fabricantes unas estimaciones sobre las ventas de sus productos al por menor a partir de muestras de tiendas minoristas. Los directores de *marketing* tienden a fijarse en la estimación y a ignorar el error de estimación. Este mes, una muestra aleatoria simple de 75 tiendas da una media de ventas de 52 unidades de un pequeño electrodoméstico, con una desviación típica de 13 unidades. Durante el mismo mes del año anterior, una muestra aleatoria simple de 53 tiendas dio unas ventas medias de 49 unidades, con una desviación típica de 11 unidades. Un aumento de 49 a 52 unidades es un incremento del 6%. El director de *marketing* está contento porque las ventas han subido un 6%.

[26] A. H. Ismail y R. J. Young, 1973, "The effect of chronic exercise on the personality of middle-aged men", *Journal of Human Ergology*, 2, págs. 47-57.

(a) Utiliza el procedimiento t de dos muestras para dar un intervalo de confianza del 95% de la diferencia entre el número medio de unidades vendidas este año y el año pasado en todas las tiendas minoristas.

(b) Explica con un lenguaje que el directivo pueda entender por qué no podemos estar seguros de que las ventas subieran un 6%, y que incluso es posible que las ventas hayan bajado.

6.44. Un banco compara dos propuestas para fomentar la utilización de las tarjetas de crédito entre sus clientes (el banco gana un porcentaje sobre la cantidad cargada en la tarjeta de crédito, que pagan las tiendas que la aceptan). La propuesta A ofrece eliminar la cuota anual para los clientes que carguen 240.000 pesetas o más durante el año. La propuesta B ofrece devolver en metálico a los clientes un pequeño porcentaje de la cantidad total cargada al final del año. El banco ofrece cada propuesta a una muestra aleatoria simple de 150 clientes con tarjeta de crédito. Al final del año el banco registra la cantidad total cargada por cada cliente en su tarjeta de crédito. He aquí los estadísticos de resumen de estas cantidades.

Grupo	n	$\bar{x}$ pesetas	s pesetas
A	150	198.700	39.200
B	150	205.600	41.300

(a) Los datos, ¿muestran una diferencia significativa entre las cantidades medias cargadas por los clientes de las dos opciones? Plantea las hipótesis nula y alternativa, y calcula el estadístico t de dos muestras. Obtén el valor P. Expón tus conclusiones prácticas.

(b) Las distribuciones de las cantidades cargadas son asimétricas hacia la derecha, pero no hay observaciones atípicas debido a los límites que impone el banco en las cantidades que se pueden cargar en las tarjetas. ¿Crees que la asimetría amenaza la validez de la prueba que utilizaste en (a)? Justifica tu respuesta.

(c) El estudio del banco, ¿es un experimento? ¿Por qué? ¿Cómo afecta esto a las conclusiones que el banco pueda extraer de este estudio?

6.45. Una maestra cree que unas nuevas actividades de lectura en clase ayudarán a mejorar la capacidad lectora de los niños de primaria. La maestra programa que 21 niños de una clase de tercero sigan estas actividades durante 8 semanas. Una clase control de 23 alumnos de tercero sigue el mismo programa académico, pero sin realizar las nuevas actividades de lectura. Al final del período de 8 semanas todos los alumnos pasan la prueba de lectura DRP (*Degree of Reading Power*), que mide aquellos aspectos de la capacidad lectora de los niños que deberían mejorar con el tratamiento establecido. He aquí los datos.[27]

Tratamiento					Control				
24	43	58	71	43	42	43	55	26	62
49	61	44	67	49	37	33	41	19	54
53	56	59	52	62	20	85	46	10	17
54	57	33	46	43	60	53	42	37	42
57					55	28	48		

(a) Examina los datos gráficamente. ¿Hay observaciones atípicas o asimetrías importantes que pudieran impedir la utilización de los procedimientos t?

(b) ¿Existe suficiente evidencia de que las nuevas actividades mejoran el resultado medio en el examen DRP? Lleva a cabo una prueba e informa de tus resultados.

(c) Aunque este estudio es un experimento, su diseño no es el ideal, porque tuvo que hacerse en la escuela sin alterar la actividad normal de las clases. ¿Qué aspecto de un buen diseño experimental no se ha tenido en cuenta?

6.46. Los investigadores que estudian el aprendizaje del habla suelen comparar mediciones hechas sobre grabaciones del habla de adultos y de niños. Una variable de interés es el momento del inicio de la voz (MIV). He aquí los resultados de niños de 6 años y de adultos a los que se les pidió que pronunciasen la palabra *"bees"*. El MIV se mide en milisegundos y puede ser positivo o negativo.[28]

Grupo	n	$\bar{x}$	s
Niños	10	− 3.67	33.89
Adultos	20	− 23.17	50.74

(a) Los investigadores querían saber si el MIV diferencia a los adultos de los niños. Plantea H_0 y H_a, y lleva a cabo una prueba t de dos muestras. Da un valor P y saca tus conclusiones.

(b) Da un intervalo de confianza del 95% de la diferencia entre las medias del MIV de niños y adultos cuando se pronuncia la palabra *"bees"*. Explica por qué sabías a partir de tu resultado en (a) que este intervalo contendría el 0 (ninguna diferencia).

[27] Maribeth Cassidy Schmitt, *The Effects of an Elaborated Directed Reading Activity on the Metacomprehension Skills of Third Graders*, tesis doctoral, Purdue University, 1987.

[28] M. A. Zlatin y R. A. Koenigsknecht, 1976, "Development of the voicing contrast: a comparison of voice onset time in stop perception and production", *Journal of Speech and Hearing Research*, 19, págs. 93-111.

6.47. Los investigadores del estudio comentado en el ejercicio 6.46 analizaron los MIV de adultos y niños al pronunciar distintas palabras. Explica por qué no deberían hacer una prueba t de dos muestras distinta para cada palabra y concluir que aquellas palabras con una diferencia significativa ($P < 0{,}05$) distinguen a los niños de los adultos (los investigadores no cometieron este error).

Los siguientes ejercicios tratan sobre la potencia de la prueba t de dos muestras, un tema optativo. Si has leído la sección 5.5 y la discusión sobre la potencia de la prueba t de una sola muestra descritas en el apartado 6.2.5, el ejercicio 6.48 te orientará sobre cómo hallar la potencia de la prueba t de dos muestras.

6.48 (Optativo). En el ejemplo 6.8, un pequeño estudio sobre hombres negros sugirió que un suplemento de calcio podía reducir la presión sanguínea. Ahora proyectamos un experimento médico de más envergadura sobre este mismo efecto. Queremos utilizar 100 sujetos en cada uno de los dos grupos. Los tamaños muestrales, ¿son suficientemente grandes para hacer muy probable que el estudio proporcione una fuerte evidencia ($\alpha = 0{,}01$) del efecto del calcio, si de hecho el calcio disminuye la presión sanguínea en 5 milímetros más que un placebo? Para contestar esta pregunta calcularemos la potencia de la prueba t de dos muestras

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

en contra de la alternativa concreta $\mu_1 - \mu_2 = 5$. Basándonos en el estudio piloto del ejemplo 6.8, tomamos 8, el mayor de los dos valores s observados, como una estimación aproximada de la σ poblacional y de la s de la futura muestra.

(a) ¿Cuál es el valor aproximado del valor crítico del estadístico t de dos muestras t^* para $\alpha = 0{,}01$, cuando $n_1 = n_2 = 100$?

(b) Paso 1. *Escribe la regla para rechazar H_0 en términos de $\bar{x}_1 - \bar{x}_2$*. La prueba rechaza H_0 cuando

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \geq t^*$$

Considera que tanto s_1 como s_2 son iguales a 8, y que n_1 y n_2 son iguales a 100. Halla el número c tal que la prueba rechace H_0 cuando $\bar{x}_1 - \bar{x}_2 \geq c$.

(c) Paso 2. *La potencia es la probabilidad de rechazar H_0 cuando la alternativa es cierta*. Supón que $\mu_1 - \mu_2 = 5$ y que tanto σ_1 como σ_2 son iguales a 8. La potencia que buscamos es la probabilidad de que $\bar{x}_1 - \bar{x}_2 \geq c$ bajo estos supuestos. Calcula dicha potencia.