Capítulo 3: Describir datos con estadísticas

### **Cálculo de la correlación entre dos conjuntos de datos**

En este apartado aprenderemos a calcular una medida estadística que nos indica la naturaleza y la fuerza de la relación entre dos conjuntos de números: el coeficiente de *correlación de Pearson*, que llamaré simplemente *coeficiente de correlación*. Observa que este coeficiente mide la fuerza de la relación *lineal*. Tendríamos que utilizar otras medidas (que no trataremos aquí) para averiguar el coeficiente cuando dos conjuntos tienen una relación no lineal. El coeficiente puede ser positivo o negativo, y su magnitud puede oscilar entre -1 y 1 (ambos inclusive).

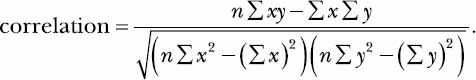
Un coeficiente de correlación de 0 indica que no hay correlación lineal entre las dos cantidades. (Ten en cuenta que esto no significa que las dos cantidades sean independientes entre sí. Podría seguir existiendo una relación no lineal entre ellas, por ejemplo). Un coeficiente de 1 o cercano a 1 indica que existe una fuerte correlación lineal positiva; un coeficiente de exactamente 1 se denomina correlación positiva perfecta. Del mismo modo, un coeficiente de correlación de -1 o cercano a -1 indica una fuerte correlación negativa, donde 1 indica una correlación negativa perfecta.

**CORRELACIÓN Y CAUSALIDAD**

En estadística, a menudo te encontrarás con la afirmación "correlación no implica causalidad". Esto es un recordatorio de que aunque dos conjuntos de observaciones estén fuertemente correlacionados entre sí, eso no significa que una variable *cause* la otra. Cuando dos variables están fuertemente correlacionadas, a veces hay un tercer factor que influye en ambas variables y explica la correlación. Un ejemplo clásico es la correlación entre las ventas de helados y los índices de delincuencia: si haces un seguimiento de estas dos variables en una ciudad típica, es probable que encuentres una correlación, pero esto no significa que las ventas de helados causen la delincuencia (o viceversa). La venta de helados y la delincuencia están correlacionadas porque ambas aumentan cuando hace más calor en verano. Por supuesto, esto tampoco significa que el calor provoque directamente el aumento de la delincuencia; también hay causas más complicadas detrás de esa correlación.

#### ***Cálculo del coeficiente de correlación***

El coeficiente de correlación se calcula mediante la fórmula



En la fórmula anterior, *n* es el número total de valores presentes en cada conjunto de números (los conjuntos tienen que tener la misma longitud). Los dos conjuntos de números se denotan por *x* e *y* (no importa cuál denotes como cuál). Los demás términos se describen como sigue

|  |  |
| --- | --- |
| *Σxy* | Suma de los productos de los elementos individuales de los dos conjuntos de números, *x* e *y* |
| *Σx* | Suma de los números del conjunto *x* |
| *Σy* | Suma de los números del conjunto *y* |
| (*Σx*)2 | Cuadrado de la suma de los números del conjunto *x* |
| (*Σy*)2 | Cuadrado de la suma de los números del conjunto *y* |
| Σx2 | Suma de los cuadrados de los números del conjunto *x* |
| Σy2 | Suma de los cuadrados de los números del conjunto *y* |

Una vez calculados estos términos, puedes combinarlos según la fórmula anterior para hallar el coeficiente de correlación. Para listas pequeñas, es posible hacerlo a mano sin demasiado esfuerzo, pero ciertamente se complica a medida que aumenta el tamaño de cada conjunto de números.

Dentro de un momento, escribiremos un programa que calcule por nosotros el coeficiente de correlación. En este programa, utilizaremos la función zip(), que nos ayudará a calcular la suma de los productos de los dos conjuntos de números. Aquí tienes un ejemplo de cómo funciona la función zip():

>>> simple\_list1 = [1, 2, 3]  
>>> simple\_list2 = [4, 5, 6]  
>>> for x, y in zip(simple\_list1, simple\_list2):  
print(x, y)  
  
1 4  
2 5  
3 6

La función zip() devuelve pares de los elementos correspondientes en x y y, que luego puedes utilizar en un bucle para realizar otras operaciones (como imprimir, como se muestra en el código anterior). Si las dos listas son desiguales en longitud, la función termina cuando se han leído todos los elementos de la lista más pequeña.

Ahora estamos listos para escribir un programa que calcule por nosotros el coeficiente de correlación:

def find\_corr\_x\_y(x,y):  
n = len(x)  
  
# Find the sum of the products  
prod = []  
➊ for xi,yi in zip(x,y):  
prod.append(xi\*yi)  
  
➋ sum\_prod\_x\_y = sum(prod)  
➌ sum\_x = sum(x)  
➍ sum\_y = sum(y)  
squared\_sum\_x = sum\_x\*\*2  
squared\_sum\_y = sum\_y\*\*2  
  
x\_square = []  
➎ for xi in x:  
x\_square.append(xi\*\*2)  
# Find the sum  
x\_square\_sum = sum(x\_square)  
  
y\_square=[]  
for yi in y:  
y\_square.append(yi\*\*2)  
# Find the sum  
y\_square\_sum = sum(y\_square)  
  
# Use formula to calculate correlation  
➏ numerator = n\*sum\_prod\_x\_y - sum\_x\*sum\_y  
denominator\_term1 = n\*x\_square\_sum - squared\_sum\_x  
denominator\_term2 = n\*y\_square\_sum - squared\_sum\_y  
➐ denominator = (denominator\_term1\*denominator\_term2)\*\*0.5  
➑ correlation = numerator/denominator  
  
return correlation

La función find\_corr\_x\_y() acepta dos argumentos, x y y, que son los dos conjuntos de números para los que queremos calcular la correlación. Al principio de esta función, hallamos la longitud de las listas y creamos una etiqueta, n, para referirnos a ella. A continuación, en ➊, tenemos un bucle for que utiliza la función zip() para calcular el producto de los valores correspondientes de cada lista (multiplicando juntos el primer elemento de cada lista, luego el segundo elemento de cada lista, y así sucesivamente). Utilizamos el método append() para añadir estos productos a la lista etiquetada prod.

En ➋, calculamos la suma de los productos almacenados en prod utilizando la función sum(). En las sentencias ➌ y ➍, calculamos la suma de los números de x y y, respectivamente (una vez más, utilizando la función sum() ). A continuación, calculamos los cuadrados de la suma de los elementos en x y y, creando las etiquetas squared\_sum\_x y squared\_sum\_y para referirnos a ellos, respectivamente.

En el bucle que comienza en ➎, calculamos el cuadrado de cada uno de los elementos de x y hallamos la suma de estos cuadrados. A continuación, hacemos lo mismo con los elementos de y. Ahora tenemos todos los términos que necesitamos para calcular la correlación, y lo hacemos en las sentencias ➏,➐ y ➑. Por último, devolvemos la correlación. La correlación es una medida muy citada en los estudios estadísticos, tanto en los medios de comunicación populares como en los artículos científicos. A veces sabemos de antemano que existe una correlación, y sólo queremos averiguar la fuerza de esa correlación. Veremos un ejemplo de esto en "[Lectura de datos de un archivo CSV](ch03.html#ch03lev2sec09)" en [la página 86](ch03.html#page_86), cuando calculemos la correlación entre los datos leídos de un archivo. Otras veces, sólo sospechamos que puede haber una correlación, y debemos investigar los datos para verificar si realmente la hay (como en el ejemplo siguiente).

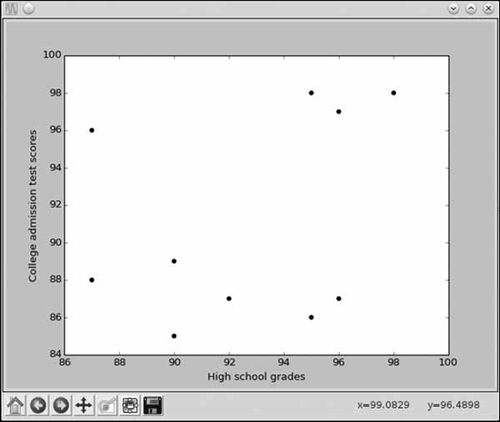
#### ***Notas de bachillerato y rendimiento en las pruebas de acceso a la universidad***

En este apartado, consideraremos un grupo ficticio de 10 estudiantes de bachillerato e investigaremos si existe una relación entre sus calificaciones en la escuela y su rendimiento en las pruebas de acceso a la universidad. [La Tabla 3-2](ch03.html#ch3tab2) enumera los datos que vamos a asumir para nuestro estudio y en los que vamos a basar nuestros experimentos. La columna "Notas de bachillerato" muestra las puntuaciones percentiles de las notas de los alumnos en el bachillerato, y la columna "Puntuaciones en el examen de acceso a la universidad" muestra sus puntuaciones percentiles en el examen de acceso a la universidad.

**Tabla 3-2:** Calificaciones de bachillerato y resultados en la prueba de acceso a la universidad

| **Notas de bachillerato** | **Puntuaciones en la prueba de acceso a la universidad** |
| --- | --- |
| 90 | 85 |
| 92 | 87 |
| 95 | 86 |
| 96 | 97 |
| 87 | 96 |
| 87 | 88 |
| 90 | 89 |
| 95 | 98 |
| 98 | 98 |
| 96 | 87 |

Para analizar estos datos, veamos un *diagrama de* dispersión. [La Figura 3-2](ch03.html#ch3fig2) muestra el diagrama de dispersión del conjunto de datos anterior, con el *eje x* representando las calificaciones de bachillerato y el *eje y* representando los resultados correspondientes de las pruebas de acceso a la universidad.



*Figura 3-2: Diagrama de dispersión de las notas de bachillerato y los resultados de las pruebas de acceso a la universidad*

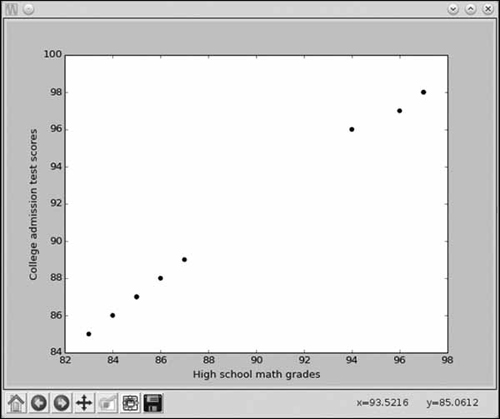
El diagrama de los datos indica que los alumnos con las notas más altas en el instituto no obtuvieron necesariamente mejores resultados en las pruebas de acceso a la universidad y viceversa. Algunos estudiantes con malas notas en el instituto obtuvieron muy buenos resultados en el examen de acceso a la universidad, mientras que otros tenían notas excelentes pero obtuvieron resultados relativamente malos en el examen de acceso a la universidad. Si calculamos el coeficiente de correlación de los dos conjuntos de datos (utilizando nuestro programa de antes), vemos que es aproximadamente 0,32. Esto significa que existe cierta correlación, pero no muy fuerte. Si la correlación fuera más cercana a 1, también lo veríamos reflejado en el diagrama de dispersión: los puntos se ajustarían más a una línea recta diagonal.

Supongamos que las notas de bachillerato que aparecen en [la Tabla 3-2](ch03.html#ch3tab2) son una media de las notas individuales en matemáticas, ciencias, inglés y ciencias sociales. Imaginemos también que el examen de la universidad hace mucho hincapié en las matemáticas, mucho más que en otras asignaturas. En lugar de fijarnos en las calificaciones generales de los alumnos en el instituto, veamos sólo sus calificaciones en matemáticas para ver si eso predice mejor cómo les fue en el examen de la universidad. [La Tabla 3-3](ch03.html#ch3tab3) muestra ahora sólo las calificaciones en matemáticas (como percentiles) y las pruebas de acceso a la universidad. El gráfico de dispersión correspondiente se muestra en [la Figura 3-3](ch03.html#ch3fig3).

**Tabla 3-3:** Notas de Matemáticas en el instituto y resultados en los exámenes de acceso a la universidad

| **Notas de matemáticas en el instituto** | **Puntuaciones en las pruebas de acceso a la universidad** |
| --- | --- |
| 83 | 85 |
| 85 | 87 |
| 84 | 86 |
| 96 | 97 |
| 94 | 96 |
| 86 | 88 |
| 87 | 89 |
| 97 | 98 |
| 97 | 98 |
| 85 | 87 |

Ahora, el diagrama de dispersión[(Figura 3-3](ch03.html#ch3fig3)) muestra que los puntos de datos se extienden casi perfectamente a lo largo de una línea recta. Esto indica que existe una alta correlación entre las puntuaciones obtenidas en matemáticas en el instituto y el rendimiento en el examen de acceso a la universidad. El coeficiente de correlación, en este caso, resulta ser aproximadamente 1. Con la ayuda del diagrama de dispersión y el coeficiente de correlación, podemos concluir que, efectivamente, existe una fuerte relación en este conjunto de datos entre las calificaciones en matemáticas del bachillerato y el rendimiento en las pruebas de acceso a la universidad.



*Figura 3-3: Diagrama de dispersión de las calificaciones en matemáticas de bachillerato y los resultados en las pruebas de acceso a la universidad*

[anterior](ch03_5.html)[Subtema 6 de 10: (Ver todo)](ch03.html)[siguiente](ch03_7.html)