Capítulo 3: Describir datos con estadísticas

### **Medir la dispersión**

Los siguientes cálculos estadísticos que veremos miden la *dispersión*, que nos indica la distancia que separa los números de un conjunto de datos de la media del conjunto. Aprenderemos a calcular tres medidas distintas de la dispersión: el rango, la varianza y la desviación típica.

#### ***Hallar el rango de un conjunto de números***

Una vez más, considera la lista de donaciones durante el periodo A: 100, 60, 70, 900, 100, 200, 500, 500, 503, 600, 1000 y 1200. Comprobamos que la media de donaciones por día es de 477,75. Pero sólo mirando la media, no tenemos ni idea de si todas las donaciones cayeron en un rango estrecho -digamos entre 400 y 500- o si variaron mucho más que eso -digamos entre 60 y 1200, como en este caso-. Para una lista de números, el *rango* es la diferencia entre el número más alto y el más bajo. Podrías tener dos grupos de números con exactamente la misma media pero con rangos muy diferentes, por lo que conocer el rango aporta más información sobre un conjunto de números que la que podemos obtener con sólo mirar la media, la mediana y la moda.

El siguiente programa halla el rango de la lista de donaciones anterior:

'''  
Find the range  
'''  
  
def find\_range(numbers):  
  
➊ lowest = min(numbers)  
➋ highest = max(numbers)  
# Find the range  
r = highest-lowest  
  
➌ return lowest, highest, r  
  
if \_\_name\_\_ == '\_\_main\_\_':  
donations = [100, 60, 70, 900, 100, 200, 500, 500, 503, 600, 1000, 1200]  
➍ lowest, highest, r = find\_range(donations)  
print('Lowest: {0} Highest: {1} Range: {2}'.format(lowest, highest, r))

La función find\_range() acepta una lista como parámetro y halla el rango. Primero, calcula los números más bajos y más altos utilizando las funciones min() y max() en ➊ y ➋. Como indican los nombres de las funciones, encuentran los valores mínimo y máximo en una lista de números.

A continuación, calculamos el rango tomando la diferencia entre los números mayor y menor, utilizando la etiqueta r para referirnos a esta diferencia. En ➌, devolvemos los tres números: el número más bajo, el número más alto y el rango. Es la primera vez en el libro que devolvemos varios valores de una función: en lugar de devolver sólo un valor, esta función devuelve tres. En ➍, utilizamos tres etiquetas para *recibir* los tres valores que devuelve la función find\_range(). Por último, imprimimos los valores. Cuando ejecutes el programa, deberías ver la siguiente salida:

Lowest: 60 Highest: 1200 Range: 1140

Esto nos indica que las donaciones totales de los días estaban bastante repartidas, con un rango de 1140, porque teníamos totales diarios tan pequeños como 60 y tan grandes como 1200.

#### ***Hallar la varianza y la desviación típica***

El rango nos indica la diferencia entre los dos extremos de un conjunto de cifras, pero ¿qué ocurre si queremos saber más sobre cómo varían todas las cifras individuales respecto a la media? ¿Eran todas similares, agrupadas cerca de la media, o eran todas diferentes, más cerca de los extremos? Hay dos medidas de dispersión relacionadas que nos dicen más sobre una lista de números en este sentido: la *varianza* y la *desviación típica*. Para calcular cualquiera de ellas, primero tenemos que hallar la diferencia de cada uno de los números respecto a la media. La varianza es la media de los cuadrados de esas diferencias. Una varianza alta significa que los valores están lejos de la media; una varianza baja significa que los valores están agrupados cerca de la media. Calculamos la varianza mediante la fórmula

image

En la fórmula, *xi* representa los números individuales (en este caso, los donativos totales diarios), xmedia representa la media de estos números (el donativo medio diario), y *n* es el número de valores de la lista (el número de días en los que se recibieron donativos). Para cada valor de la lista, tomamos la diferencia entre ese número y la media y la elevamos al cuadrado. Luego, sumamos todas esas diferencias al cuadrado y, por último, dividimos la suma total por *n* para hallar la varianza.

Si queremos calcular también la desviación típica, basta con sacar la raíz cuadrada de la varianza. Los valores que están dentro de una desviación típica de la media se pueden considerar bastante típicos, mientras que los valores que están a tres o más desviaciones típicas de la media se pueden considerar mucho más atípicos; a esos valores los llamamos *valores atípicos*.

¿Por qué tenemos estas dos medidas de dispersión: la varianza y la desviación típica? En resumen, las dos medidas son útiles en situaciones diferentes. Volviendo a la fórmula que utilizamos para calcular la varianza, puedes ver que la varianza se expresa en unidades cuadradas porque es la media de la diferencia al cuadrado respecto a la media. Para algunas fórmulas matemáticas, es más agradable trabajar con esas unidades cuadradas en lugar de sacar la raíz cuadrada para hallar la desviación típica. Por otra parte, la desviación típica se expresa en las mismas unidades que los datos de la población. Por ejemplo, si calculas la varianza de nuestra lista de donaciones (como haremos dentro de un momento), el resultado se expresa en dólares al cuadrado, lo que no tiene mucho sentido. Mientras tanto, la desviación típica se expresa simplemente en dólares, la misma unidad que cada una de las donaciones.

El siguiente programa halla la varianza y la desviación típica de una lista de números:

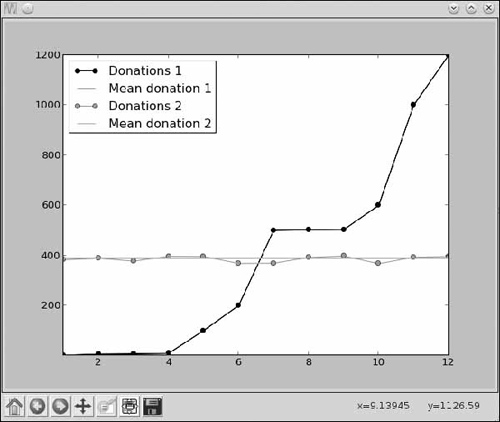
'''  
Find the variance and standard deviation of a list of numbers  
'''  
  
def calculate\_mean(numbers):  
s = sum(numbers)  
N = len(numbers)  
# Calculate the mean  
mean = s/N  
  
return mean  
  
def find\_differences(numbers):  
# Find the mean  
mean = calculate\_mean(numbers)  
# Find the differences from the mean  
diff = []  
for num in numbers:  
diff.append(num-mean)  
  
return diff  
  
def calculate\_variance(numbers):  
  
# Find the list of differences  
➊ diff = find\_differences(numbers)  
# Find the squared differences  
squared\_diff = []  
➋ for d in diff:  
squared\_diff.append(d\*\*2)  
# Find the variance  
sum\_squared\_diff = sum(squared\_diff)  
➌ variance = sum\_squared\_diff/len(numbers)  
return variance  
  
if \_\_name\_\_ == '\_\_main\_\_':  
donations = [100, 60, 70, 900, 100, 200, 500, 500, 503, 600, 1000, 1200]  
variance = calculate\_variance(donations)  
print('The variance of the list of numbers is {0}'.format(variance))  
  
➍ std = variance\*\*0.5  
print('The standard deviation of the list of numbers is {0}'.format(std))

La función calculate\_variance() calcula la varianza de la lista de números que se le pasa. Primero, llama a la función find\_differences() en ➊ para calcular la diferencia de cada uno de los números respecto a la media. La función find\_differences() devuelve la diferencia de cada donación respecto al valor medio en forma de lista. En esta función, utilizamos la función calculate\_mean() que escribimos antes para hallar la donación media. A continuación, a partir de ➋, se calculan los cuadrados de estas diferencias y se guardan en una lista denominada squared\_diff. A continuación, utilizamos la función sum() para hallar la suma de las diferencias al cuadrado y, por último, calculamos la varianza en ➌. En ➍, calculamos la desviación típica tomando la raíz cuadrada de la varianza.

Cuando ejecutes el programa anterior, deberías ver el siguiente resultado:

The variance of the list of numbers is 141047.35416666666  
The standard deviation of the list of numbers is 375.5627166887931

Tanto la varianza como la desviación típica son muy grandes, lo que significa que las donaciones totales diarias individuales varían mucho respecto a la media. Ahora, comparemos la varianza y la desviación típica de un conjunto diferente de donaciones que tienen la misma media: 382, 389, 377, 397, 396, 368, 369, 392, 398, 367, 393 y 396. En este caso, la varianza y la desviación típica resultan ser 135,38888888888889 y 11,63567311713804, respectivamente. Los valores más bajos de la varianza y la desviación típica nos indican que los números individuales están más cerca de la media. [La figura 3-1](ch03.html#ch3fig1) ilustra visualmente este punto.



*Figura 3-1: Variación de los donativos en torno al donativo medio*

Las donaciones medias de ambas listas de donaciones son similares, por lo que las dos líneas se superponen, apareciendo como una sola línea en la figura. Sin embargo, los donativos de la primera lista varían mucho respecto a la media, mientras que los donativos de la segunda lista están muy cerca de la media, lo que confirma lo que dedujimos del valor más bajo de la varianza.

[anterior](ch03_4.html)[Subtema 5 de 10: (Ver todo)](ch03.html)[siguiente](ch03_6.html)