Capítulo 3: Describir datos con estadísticas

### **Gráficos de dispersión**

En el apartado anterior, vimos un ejemplo de cómo un diagrama de dispersión puede darnos una primera indicación de la existencia de una correlación entre dos conjuntos de números. En este apartado, veremos la importancia de analizar gráficos de dispersión observando un conjunto de cuatro conjuntos de datos. Para estos conjuntos de datos, las medidas estadísticas convencionales resultan ser todas iguales, pero los gráficos de dispersión de cada conjunto de datos revelan importantes diferencias.

En primer lugar, vamos a ver cómo crear un gráfico de dispersión en Python:

>>> x = [1, 2, 3, 4]  
>>> y = [2, 4, 6, 8]  
>>> import matplotlib.pyplot as plt  
➊ >>> plt.scatter(x, y)  
<matplotlib.collections.PathCollection object at 0x7f351825d550>  
>>> plt.show()

La función scatter() se utiliza para crear un gráfico de dispersión entre dos listas de números, x y y ➊. La única diferencia entre este gráfico y los gráficos que creamos en [el Capítulo 2](ch02.html#ch02) es que aquí utilizamos la función scatter() en lugar de la función plot(). Una vez más, tenemos que llamar a show() para visualizar el gráfico.

Para aprender más sobre los gráficos de dispersión, veamos un importante estudio estadístico: "Los gráficos en el análisis estadístico", del estadístico Francis Anscombe[.](footnote.html#fn01) 1 El estudio considera cuatro conjuntos de datos diferentes -denominados *el cuarteto de Anscombe- con*idénticas propiedades estadísticas: media, varianza y coeficiente de correlación.

Los conjuntos de datos son los que se muestran en [la Tabla 3-4](ch03.html#ch3tab4) (reproducida del estudio original).

**Tabla 3-4:** Cuarteto de Anscombe: cuatro conjuntos de datos diferentes con medidas estadísticas casi idénticas

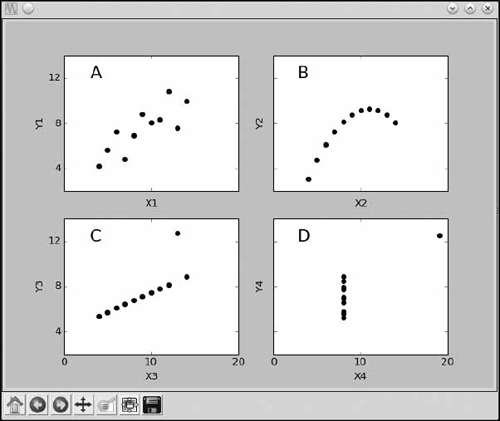
| **A** | | **B** | | **C** | | **D** | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **Y1** | **X2** | **Y2** | **X3** | **Y3** | **X4** | **Y4** |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

Nos referiremos a los pares (X1, Y1), (X2, Y2), (X3, Y3) y (X4, Y4) como conjuntos de datos A, B, C y D, respectivamente. [La Tabla 3-5](ch03.html#ch3tab5) presenta las medidas estadísticas de los conjuntos de datos redondeadas a dos cifras decimales.

**Tabla 3-5:** Cuarteto de Anscombe-Medidas estadísticas

| **Conjunto de datos** | **X** | | **Y** | |  |
| --- | --- | --- | --- | --- | --- |
| **Media** | **Desv. est.** | **Media** | **Desv. est.** | **Correlación** |
| A | 9.00 | 3.32 | 7.50 | 2.03 | 0.82 |
| B | 9.00 | 3.32 | 7.50 | 2.03 | 0.82 |
| C | 9.00 | 3.32 | 7.50 | 2.03 | 0.82 |
| D | 9.00 | 3.32 | 7.50 | 2.03 | 0.82 |

Los diagramas de dispersión de cada conjunto de datos se muestran en [la Figura](ch03.html#ch3fig4) 3-4.



*Figura 3-4: Diagramas de dispersión del cuarteto de Anscombe*

Si nos fijamos sólo en las medidas estadísticas tradicionales (ver [Tabla 3-5](ch03.html#ch3tab5)) -como la media, la desviación típica y el coeficiente de correlación-, estos conjuntos de datos parecen casi idénticos. Pero los gráficos de dispersión muestran que estos conjuntos de datos son en realidad muy diferentes entre sí. Así pues, los gráficos de dispersión pueden ser una herramienta importante y deben utilizarse junto con otras medidas estadísticas antes de sacar conclusiones sobre un conjunto de datos.

[anterior](ch03_6.html)[Subtema 7 de 10: (Ver todo)](ch03.html)[siguiente](ch03_8.html)