Comparemos conjuntos de datos

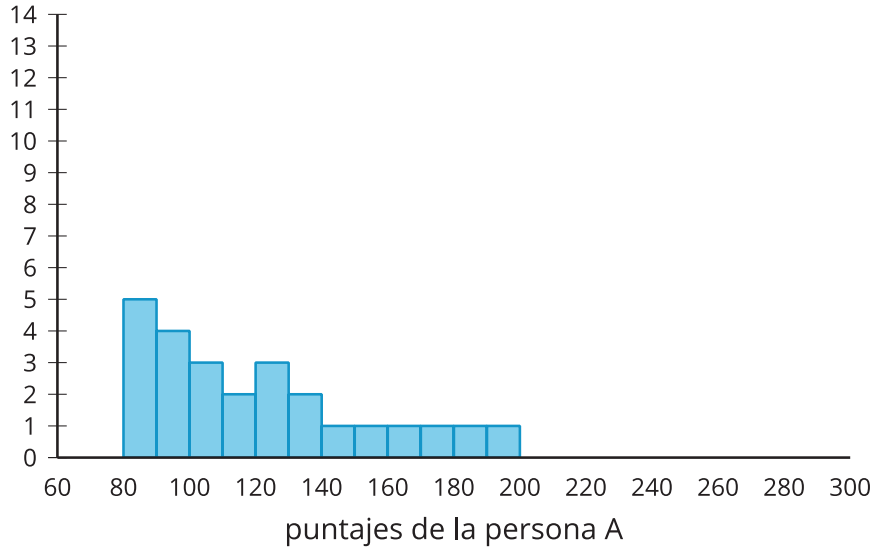
Comparemos estadísticos de conjuntos de datos.

15.1 Compañeros de bolos

En cada histograma se muestran las distribuciones de los puntajes de una persona distinta en los últimos 25 juegos de bolos que jugó. Escoge 2 de estas personas para que se unan a tu equipo de bolos. Explica tu razonamiento.

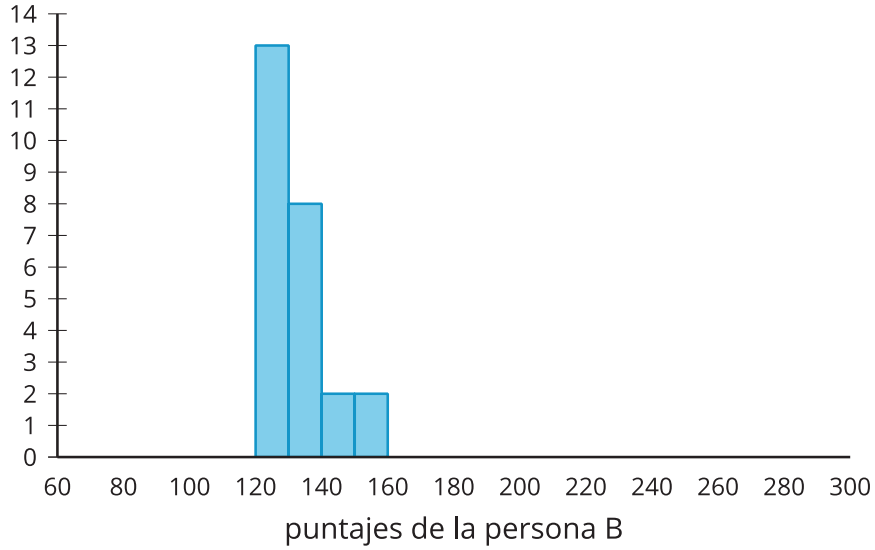
Persona A

- media: 118.96
- mediana: 111
- desviación estándar: 32.96
- rango intercuartil: 44



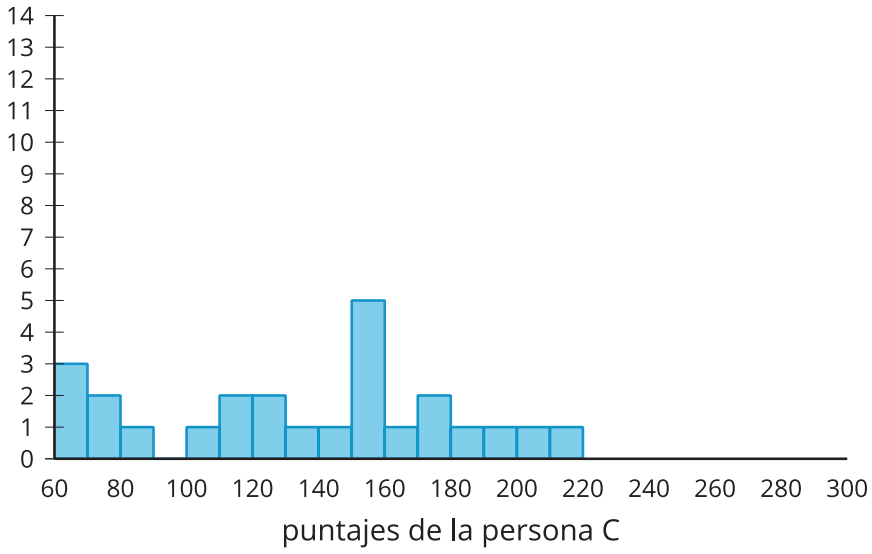
Persona B

- media: 131.08
- mediana: 129
- desviación estándar: 8.64
- rango intercuartil: 8



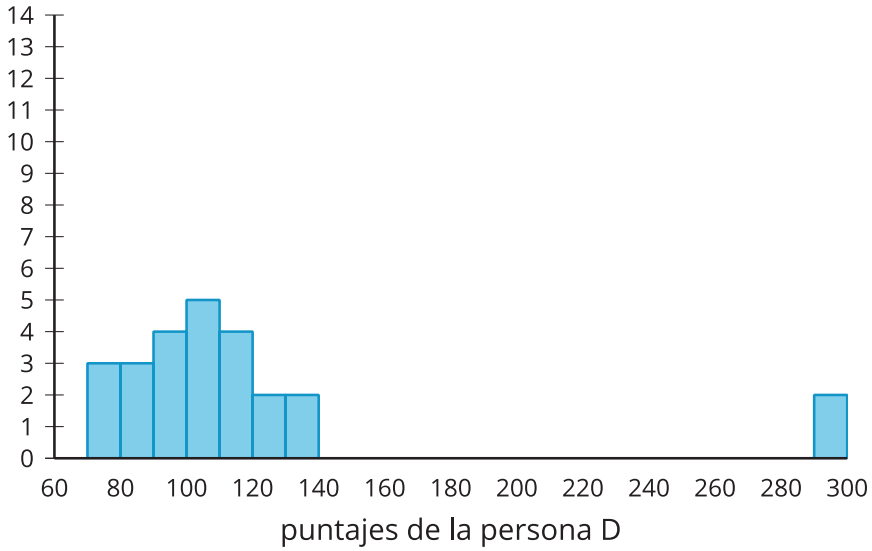
Persona C

- media: 133.92
- mediana: 145
- desviación estándar: 45.04
- rango intercuartil: 74



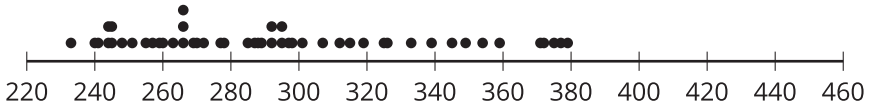
Persona D

- media: 116.56
- mediana: 103
- desviación estándar: 56.22
- rango intercuartil: 31.5

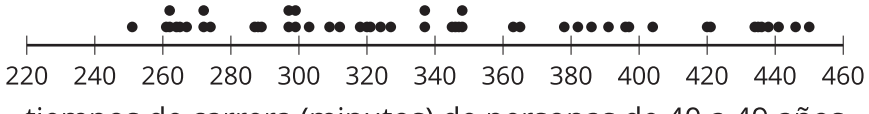


15.2 Comparemos tiempos de carrera

Se midieron los tiempos de carrera de todos los corredores de maratón de dos grupos de edades distintos. Cada diagrama de puntos representa los tiempos de carrera de uno de los grupos.



tiempos de carrera (minutos) de personas de 30 a 39 años



tiempos de carrera (minutos) de personas de 40 a 49 años

1. ¿Cuál de los dos grupos tiende a tardar más en correr la maratón? Explica tu razonamiento.

2. ¿Cuál de los dos grupos tiene tiempos más variables? Explica tu razonamiento.

¿Estás listo para más?

1. Si compararas los tiempos de carrera de un grupo de personas de 20 a 29 años con los tiempos de las dos distribuciones anteriores, ¿qué crees que observarías?

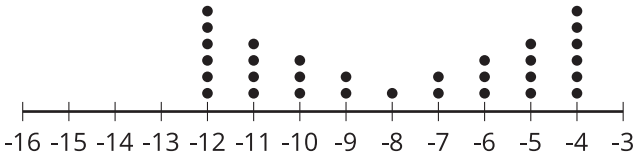
2. Encuentra algunos tiempos de carrera reales de personas en este grupo de edad y haz un diagrama de puntos o un diagrama de caja de tus datos que te ayude a compararlos con los otros.

15.3 Comparemos medidas

Para cada grupo de conjuntos de datos:

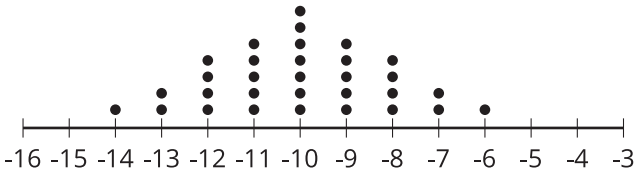
- Determina las mejores medidas de centro y de variabilidad para analizar los conjuntos basándote en la forma de cada distribución.
- Determina cuál conjunto de datos tiene la mayor medida de centro.
- Determina cuál conjunto de datos tiene la mayor medida de variabilidad.
- Prepárate para explicar tu razonamiento.

1a



distribución 1a

1b



distribución 1b

2a

distribución 2a

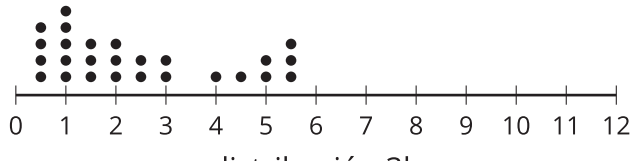
2b

distribución 2b

3a

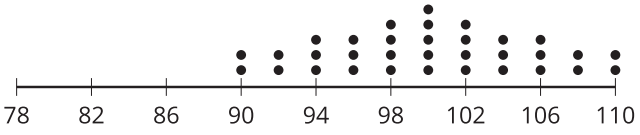
distribución 3a

3b



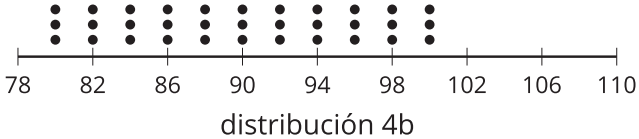
distribución 3b

4a



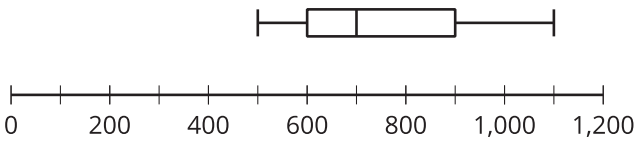
distribución 4a

4b



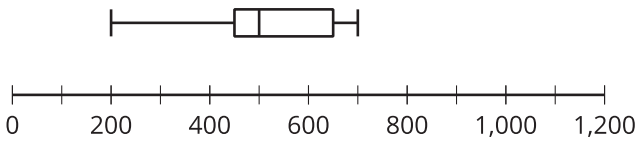
distribución 4b

5a



distribución 5a

5b



distribución 5b

6a

Un pódcast de política recibe en su mayoría comentarios de personas que lo aman o lo odian.

6b

Un pódcast de cocina recibe comentarios que son de personas que ni lo odian ni lo aman.

7a

En una prueba de resistencia del hormigón en la obra A, todas las 12 muestras se rompieron a 450 libras por pulgada cuadrada (psi, por su sigla en inglés).

7b

En una prueba de resistencia del hormigón en la obra B, las muestras se rompieron cada 10 psi: la primera capa se rompió a 450 psi y la última capa se rompió a 560 psi.

7c

En una prueba de resistencia del hormigón en la obra C, 6 muestras se rompieron a 430 psi y otras 6 se rompieron a 460 psi.

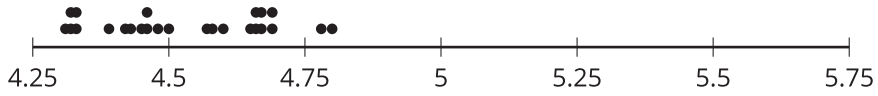
Resumen de la lección 15

Para comparar conjuntos de datos, conviene examinar sus medidas de centro y sus medidas de variabilidad. La forma de la distribución puede ayudarnos a elegir la medida de centro y la medida de variabilidad más útiles.

Cuando las distribuciones son simétricas o aproximadamente simétricas, preferimos usar la media como la medida de centro y se debe usar junto con la desviación estándar, medida preferida de variabilidad en esos casos. Cuando las distribuciones son asimétricas o cuando hay datos atípicos, la mediana es con frecuencia una mejor medida de centro y se debe usar junto con el rango intercuartil (IQR), medida preferida de variabilidad.

Después de seleccionar las medidas apropiadas de centro y de variabilidad de un conjunto de datos, estas medidas se pueden comparar con las medidas de otro conjunto, si ambos conjuntos tienen una forma similar.

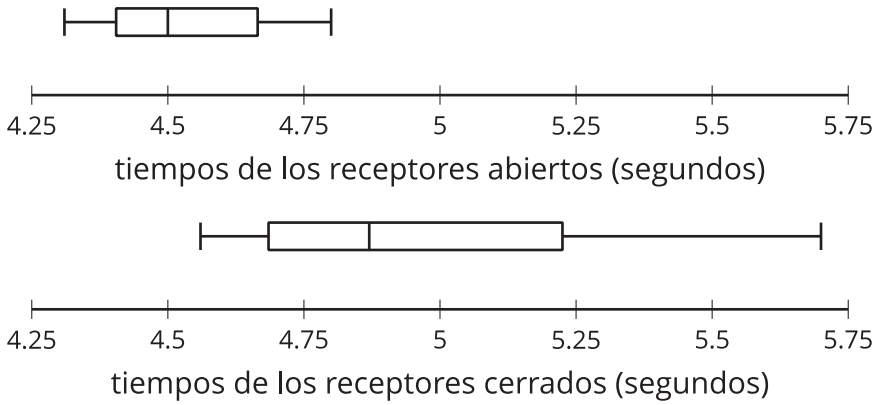
Por ejemplo, comparemos el número de segundos que tarda un jugador de fútbol americano en completar una carrera de 40 yardas en dos posiciones distintas. Primero, podemos examinar un diagrama de puntos de los datos y ver que los tiempos de los receptores cerrados no parecen estar distribuidos simétricamente, así que es probable que debamos encontrar la mediana y el IQR de ambos conjuntos de datos para comparar la información.



tiempos de los receptores abiertos (segundos)

tiempos de los receptores cerrados (segundos)

La mediana y el IQR se pueden calcular a partir de los valores, pero también se pueden determinar a partir de un diagrama de caja.



Esto muestra que los tiempos de los receptores cerrados tienen una mediana mayor (aproximadamente 4.9 segundos) comparada con la mediana de los tiempos de los receptores abiertos (aproximadamente 4.5 segundos). El IQR también es mayor para los tiempos de los receptores cerrados (aproximadamente 0.5 segundos) comparado con el IQR de los tiempos de los receptores abiertos (aproximadamente 0.25 segundos).

Esto significa que los receptores cerrados tienden a ser más lentos en la carrera de 40 yardas en comparación con los receptores abiertos. Los receptores cerrados también tienen mayor variabilidad en sus tiempos. Si consideramos todo esto, se puede interpretar que, en general, un receptor abierto típico es más rápido que un receptor cerrado típico y los receptores abiertos tienden a tener tiempos más parecidos entre sí que los receptores cerrados.