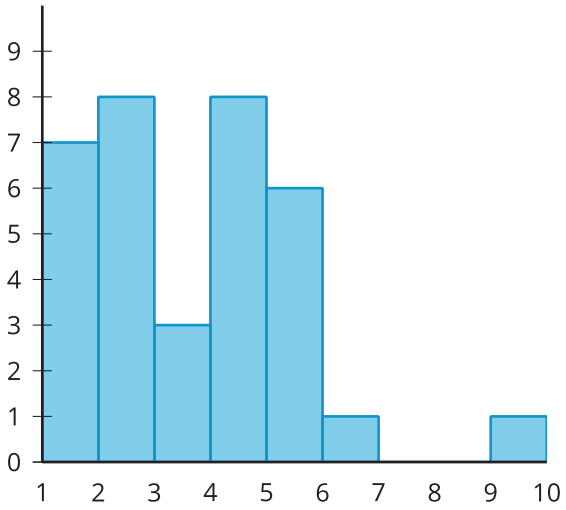
Datos atípicos

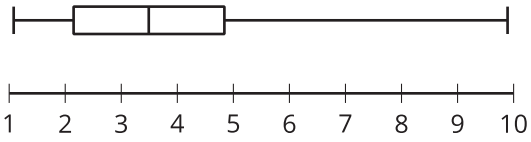
Investiguemos datos atípicos y aprendamos qué hacer con ellos.

14.1 Gasto en salud

El histograma y el diagrama de caja muestran la cantidad promedio de dinero, en miles de dólares, que cada uno de 34 países gasta en salud por cada persona (el gasto per cápita).



gasto per cápita de varios países en salud (miles de dólares)



gasto per cápita de varios países en salud (miles de dólares)

1. Hay un valor del conjunto de datos que es un **dato atípico**. ¿Cuál es? ¿Cuál es su valor aproximado?

2. Una de las reglas de decisión dice que un valor es un dato atípico si es mayor que Q3 en más de 1.5 veces el rango intercuartil. Muestra en el diagrama de caja si tu valor cumple o no con esta definición de dato atípico.

14.2 Investiguemos datos atípicos

Este es el conjunto de datos que se usó para crear el histograma y el diagrama de caja del calentamiento.

1.0803	1.0875	1.4663	1.7978	1.9702	1.9770	1.9890	2.1011	2.1495	2.2230
2.5443	2.7288	2.7344	2.8223	2.8348	3.2484	3.3912	3.5896	4.0334	4.1925
4.3763	4.5193	4.6004	4.7081	4.7528	4.8398	5.2050	5.2273	5.3854	5.4875
5.5284	5.5506	6.6475	9.8923						

1. Usa tecnología para encontrar la media, la desviación estándar y el resumen de cinco números.

2. El valor máximo de este conjunto de datos representa el gasto en salud per cápita en los Estados Unidos. ¿Este gasto debe considerarse un dato atípico? Explica tu razonamiento.

3. Aunque los datos atípicos no se deben quitar sin haber considerado su origen, es importante ver cómo estos pueden influir en varios estadísticos. Para llevar a cabo este análisis, quita el valor del gasto en los Estados Unidos del conjunto de datos.

   a. Con tecnología, calcula la media, la desviación estándar y el resumen de cinco números del nuevo conjunto de datos.

   b. Considera la media, la desviación estándar, la mediana y el rango intercuartil del conjunto de datos sin el dato atípico. Compáralos con los mismos estadísticos de resumen del conjunto de datos original. ¿Qué puedes decir?

14.3 Orígenes de los datos atípicos

1. Se ha recopilado el número de vehículos eléctricos registrados en los 39 condados del estado de Washington.
   - media: 3,589.3 automóviles
   - mínimo: 3 automóviles
   - Q1: 170 automóviles
   - mediana: 506 automóviles
   - Q3: 1,560 automóviles
   - máximo: 73,996 automóviles

   16223 337 460 60 467 73996 8368 1556 222 806 1736 3 238

   3444 165 706 3497 10657 37 278 186 424 170 45 4425 4601

   36 677 554 25 858 12 796 773 1560 44 194 841 506

   a. ¿Algunos de los valores son datos atípicos? Explica o muestra tu razonamiento.

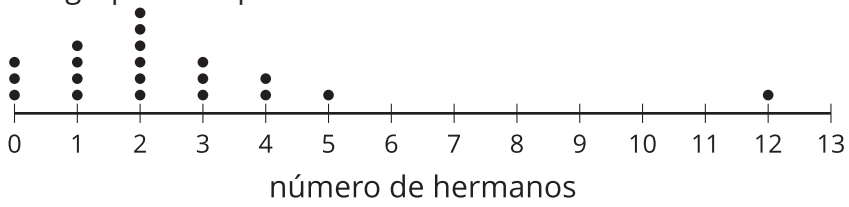
   b. Si hay datos atípicos, ¿por qué crees que pueden existir? ¿Se deben incluir en un análisis de los datos?

2. Cada una de las situaciones que se describen aquí tiene un dato atípico. En cada situación, ¿cómo decidirías si es apropiado mantener o quitar el dato atípico cuando se analicen los datos? Discute con tu compañero lo que pensaste.

   a. Un dado numérico tiene sus caras marcadas del 1 al 6. Tyler anota los resultados que obtiene al lanzar 15 veces el dado:

      1 1 1 1 2 2 3 3 4 4 5 5 5 6 20

   b. El diagrama de puntos representa la distribución del número de hermanos de los integrantes de un grupo de 20 personas.



   c. En una clase de Ciencias, 11 grupos de estudiantes están sintetizando biodiésel. Al final del experimento, cada grupo de estudiantes registró la masa, en gramos, del biodiésel que sintetizó. Las masas son:

      0 1.245 1.292 1.375 1.383 1.412 1.435 1.471 1.482 1.501

      1.532

¿Estás listo para más?

Revisa algunos de los datos numéricos que tú y tus compañeros recolectaron en la primera lección de esta unidad.

1. ¿Algunos de los valores son datos atípicos? Explica o muestra tu razonamiento.

2. Si hay datos atípicos, ¿por qué crees que pueden existir? ¿Se deben incluir en un análisis de los datos?

Resumen de la lección 14

En estadística, un **dato atípico** es un valor que es inusual porque se diferencia bastante de los otros valores del conjunto de datos.

En los conjuntos de datos puede haber datos atípicos por varias razones, incluidas, entre otras:

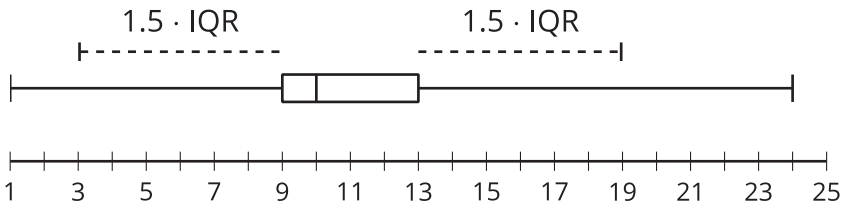
- Errores en los datos, que suceden durante de los procesos de recolección o de ingreso de los datos.
- Resultados en los datos que representan valores inusuales que ocurren en la población.

Analizar datos atípicos nos puede ayudar a descubrir casos que vale la pena estudiar en detalle o errores en el proceso de recolección de datos. En general, los datos atípicos deben ser parte de todo análisis que se realice con los datos.

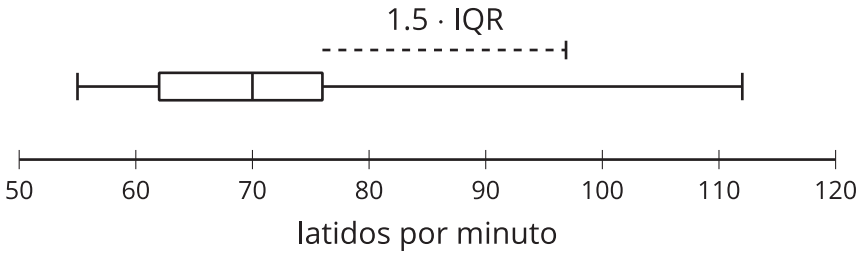
Un valor x es un dato atípico si alguna de estas condiciones ocurre:

- x es mayor que Q3 en más de 1.5 veces el rango intercuartil (si $x > \text{Q3} + 1.5 \cdot \text{IQR}$).
- x es menor que Q1 en más de 1.5 veces el rango intercuartil (si $x < \text{Q1} - 1.5 \cdot \text{IQR}$).

En este diagrama de caja hay por lo menos dos datos atípicos: el mínimo y el máximo.



Es importante identificar el origen de los datos atípicos porque estos pueden influir de manera significativa en las medidas de centro y de variabilidad. El siguiente diagrama de caja resume las frecuencias cardíacas en reposo de 50 deportistas cinco minutos después de un entrenamiento, en latidos por minuto (bpm por su sigla en inglés).



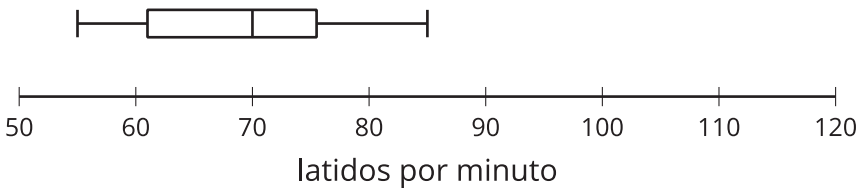
Estos son algunos estadísticos de resumen:

- media: 69.78 bpm
- desviación estándar: 10.71 bpm
- mínimo: 55 bpm
- Q1: 62 bpm
- mediana: 70 bpm
- Q3: 76 bpm
- máximo: 112 bpm

El valor máximo, 112 bpm, parece ser un dato atípico. Como el rango intercuartil es 14 bpm ($76 - 62 = 14$) y $\text{Q3} + 1.5 \cdot \text{IQR} = 97$, debemos considerar el valor máximo como un dato atípico. Al revisar todos los valores del conjunto de datos, se pudo confirmar que, en efecto, este era el único dato atípico.

Después de revisar el proceso de recolección de datos, se descubrió que la frecuencia cardíaca de 112 bpm se le midió a un deportista un minuto después del entrenamiento, en vez de cinco minutos después. El dato atípico debe borrarse del conjunto de datos porque no se obtuvo bajo las condiciones correctas.

Después de quitar el dato atípico, el diagrama de caja y los estadísticos de resumen son:



- media: 68.92 bpm
- desviación estándar: 8.9 bpm
- mínimo: 55 bpm
- Q1: 61 bpm
- mediana: 70 bpm
- Q3: 75.5 bpm
- máximo: 85 bpm

La media disminuyó 0.86 bpm y la mediana se mantuvo igual. La desviación estándar disminuyó 1.81 bpm, aproximadamente el 17% de su valor anterior. Basándose en la desviación estándar, el conjunto de datos sin el dato atípico muestra mucha menos variabilidad que el conjunto original de datos, que incluía al dato atípico. Como la media y la desviación estándar tienen en cuenta todos los valores numéricos, quitar un punto de dato muy grande puede influir ampliamente en estos estadísticos.

La mediana se mantuvo igual después de quitar el dato atípico y el IQR aumentó ligeramente. Estas medidas de centro y de variabilidad son mucho más resistentes al cambio que la media y la desviación estándar. La mediana y el IQR miden los datos de la mitad central basándose más en la cantidad de valores que en los valores numéricos en sí mismos. Así que, por lo general, la pérdida de un solo valor no tendrá un efecto tan grande en estos estadísticos.

Siempre se debe investigar el origen de cualquier posible error. Supongamos que se descubre que la medida de 112 latidos por minuto se midió bajo las condiciones correctas y que simplemente se debió a que la frecuencia cardíaca de un deportista no se redujo como la frecuencia cardíaca de los otros deportistas. En este caso, para que los datos reflejen las medidas reales, este dato no se debe borrar. Si no es posible volver a la situación para determinar el origen de un dato atípico, este no se debe quitar. Para evitar la alteración de los datos y para reportar resultados precisos, los valores de los datos no se deben borrar, a menos que se pueda confirmar que provienen de un error durante los procesos de recolección o de ingreso de los datos.