Problemas de su no tratamiento.

1. **Impacto en la estadística descriptiva:**

   - **Media y mediana:** Los outliers pueden afectar significativamente la media, ya que esta medida es sensible a valores extremos. La mediana es menos sensible, pero aún puede cambiar si hay outliers extremos.

     Si te pregunto qué medida describe mejor un conjunto de datos, ¿qué responderías?

     La mayoría de las personas diría rápidamente… ¡la media!, pero veamos si es la respuesta correcta. Supongamos que tenemos los siguientes datos:

     10,10,11,12,12,13,14,15,15,15,16,18,19

     Calculamos su media, que es 13.84615.¿Pero qué pasa si cambiamos un dato? cambiemos el último número.

     10,10,11,12,12,13,14,15,15,15,16,18,200

     Volvemos a calcular la media, que ahora es 27.77, bastante superior a la anterior.

     Con tan solo mover un valor lejos del resto, **¡la media lo seguirá!**.

     Este ejemplo indica que la **robustez** (¡vaya palabro!) del estimador es importante cuando tenemos datos atípicos (u *outliers*) y también cuando no queremos que un dato tenga más influencia que los demás en los cálculos.

     Los datos atípicos "pesan más" que los datos cercanos a la media.

     No considerar un dato extremo tiene mayores consecuencias en la estimación de la media que eliminar un datos de la región con mayor densidad.

     ¡Un solo valor es suficiente para influir enormemente en la media del conjunto de datos!

Uso de la mediana y otras variables robustas como alternativa a la media

Si calculamos la **mediana** (el valor central de una muestra ordenada) para el segundo conjuntos de datos tenemos un valor de 14 (el mismo que para el primer conjunto de datos). Vemos que este estadístico de centralidad no se ha visto perturbado por la presencia de un valor extremo, por lo tanto, es más robusto.

Veamos otras alternativas...

La **media recortada (trimming)** "desecha" los valores extremos. Es decir, elimina del análisis una fracción de los datos extremos (e.g. 20%) y calcula la media del nuevo conjunto de datos. La media recortada para nuestro caso valdría 13.67.

La **media winsorizada** progresivamente reemplaza un porcentaje de los valores extremos (e.g. 20%) por otros menos extremos.
En nuestro caso, la media winsorizada de la segunda muestra sería la misma 13.62.

Vemos que todas estas estimaciones robustas representan mejor a la muestra y se ven menos afectadas por los datos extremos.

Estadístico	Descripción	Resultado
Media aritmética	Es el valor obtenido al sumar todos los datos y dividir el resultado entre el número total de datos.	27,77
Mediana	Ordena los datos de menor a mayor y retiene el valor que divide a la distribución en dos partes iguales (i.e. igual número de datos a cada lado).	14
Media recortada	Eliminan un porcentaje (20%) de valores atípicos, y calcula la media aritmética de las observaciones restantes.	13,67
Media winsorizada	Sustituye un porcentaje (20%) de valores atípicos por el valor más próximo que no ha sido sustituido, y calcula la media aritmética del nuevo conjunto de datos.	13,62

Resultados de los estadísticos (o estimadores) que resumen el valor central para el conjunto de datos {10, 10, 11, 12, 12, 13, 14, 15, 15, 15, 16, 18, 200}.

- **Desviación estándar:** Los outliers pueden aumentar la varianza y, por lo tanto, la desviación estándar, lo que puede distorsionar la interpretación de la dispersión de los datos.

2. **Influencia en la inferencia estadística:**
    - **Intervalos de confianza:** Los outliers pueden afectar la precisión de los intervalos de confianza y sesgar la estimación de parámetros poblacionales.

- **Pruebas de hipótesis:** La presencia de outliers puede afectar la validez de las pruebas de hipótesis paramétricas, ya que muchas de estas pruebas asumen distribuciones normales.

3. **Modelos estadísticos y de machine learning:**
    - **Regresión:** Los outliers pueden tener una influencia desproporcionada en los modelos de regresión, afectando los coeficientes y la capacidad predictiva del modelo.
    - **Algoritmos sensibles a valores extremos:** Algunos algoritmos de machine learning, como la máquina de vectores de soporte (SVM) y los k-means, son sensibles a los outliers y pueden producir modelos menos robustos.
4. **Distorsión de la interpretación de los datos:**
    - Los outliers pueden llevar a interpretaciones erróneas sobre la tendencia central y la variabilidad de los datos, lo que podría afectar las decisiones basadas en esos datos.
5. **Problemas en el análisis exploratorio:**
    - En el análisis exploratorio de datos, la presencia de outliers puede distorsionar la visualización de los patrones y relaciones en los datos.
6. **Problemas en estudios epidemiológicos:**
    - En estudios de salud pública, la presencia de valores atípicos puede afectar las estimaciones de la prevalencia y la incidencia de enfermedades.
7. **Problemas en estudios financieros:**
    - En análisis financiero, la falta de tratamiento de outliers puede afectar la evaluación de riesgos y rendimientos, lo que podría conducir a decisiones financieras erróneas.

Para abordar estos problemas, es importante identificar y tratar adecuadamente los outliers mediante técnicas como la transformación de datos, la eliminación de valores atípicos o el uso de métodos robustos en el análisis estadístico y de machine learning. La elección del método dependerá del contexto específico del problema y de los datos disponibles.