Estadísticas para el clínico

# Análisis de los datos: detección de *outliers*

**Dr. Masami Yamamoto**
Clínica Universidad de Los Andes

Contacto: myamamoto@clinicauandes.cl

## Introducción

Es una necesidad del investigador clínico analizar los datos, entregar medias y calcular desviaciones estándar, para describir la población en estudio. La detección de "outliers" es parte de la estadística descriptiva. Las glicemias de un grupo de pacientes obesos previas a la cirugía bariátrica tienen una media, mediana y distribución normal en torno a ella. A pesar de esto, al evaluar un grupo de pacientes, podemos tener resultados muy alejados de la media que nos hacen sospechar si el dato de un paciente en particular es anormal para la población, y por lo tanto, excluible. Esta definición de "dato excluible" por lo extremadamente "anormal" ha sido denominada en estadística y en investigación clínica como "outlier". No hay una palabra exacta en español y se ha utilizado la palabra en inglés en muchos artículos científicos. Se ha usado el nombre "datos atípicos" en el conocido programa SPSS.

La exclusión de datos puede alterar los resultados del estudio, sus conclusiones y la aplicación clínica. Por este motivo, el criterio para excluir los "outliers" es determinante. El siguiente ejemplo muestra las glicemias preoperatorias de 30 sujetos obesos que serán sometidos a una gastrectomía parcial, para el tratamiento de su obesidad.

| | | |
|---|---|---|
| 99 | 104 | 115 |
| 123 | 75 | 116 |
| 140 | **207** | 87 |
| 87 | 65 | 78 |
| 132 | 56 | 87 |
| 75 | 101 | 94 |
| 99 | 105 | 78 |
| **188** | 68 | 89 |
| 134 | 108 | 122 |
| 122 | 121 | 134 |

El análisis descriptivo de estos datos incluyendo los casos con glicemia de 207 y 188, y luego excluyéndolos progresivamente, muestra el siguiente efecto en la media, mediana y desviación estándar:

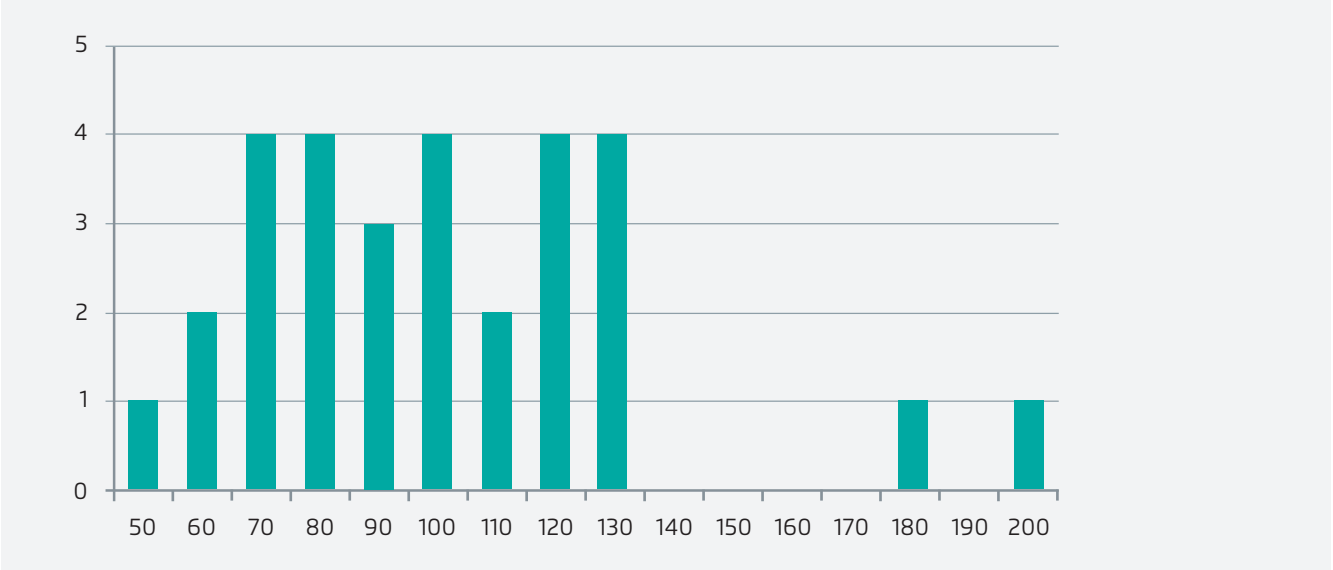
| Casos | Todos | Excluyendo 207 | Excluyendo 207 y 188 |
|---|---|---|---|
| N | 30 | 29 | 28 |
| Promedio | 107,0 | 103,5 | 100,5 |
| Mediana | 102,5 | 101,0 | 100,0 |
| DS | 33,4 | 28,0 | 23,2 |

Como se puede observar, la exclusión de dos casos, que tienen una glicemia de 207 y 188, hace que el promedio de la población baje de 107mg/dL a 100,5mg/dL y una reducción de la desviación estándar de 33,4 a 23,2. La gran interrogante es cómo decidir si excluir algunos casos considerables como *outliers*.

## Identificación de un *outlier*

En primer lugar, una observación puede ser mal interpretada como un *outlier* si la distribución de la población no es normal, como se asume en la gran mayoría de los casos. Para este punto, conviene evaluar la distribución de la población antes de excluir observaciones. La primera estrategia es realizar un histograma, es decir, un gráfico de las frecuencias relativas de las observaciones. En el caso del ejemplo, hay 30 observaciones, con pocas repeticiones de valores entre ellos, por lo que se pueden agrupar en rangos de observaciones cada 10mg/dL.

**Histograma**. Se representa el número de casos según rango de glicemias. Se observa que la mayoría de los casos está en torno a los 80-100mg/dL. Solo la visualización de este gráfico permite sospechar que la mayoría de los casos están en torno a la media, y que hay dos observaciones posibles de ser "*outliers*".




## Teoría de la detección de *outliers*

Una forma simple para calcular *outliers*, es definir el límite como 1,5 veces el rango intercuartil, por sobre el tercer cuartil o bajo el primero [1]. En 1950, Grubb publicó un método para detectar *outliers* únicos en una serie de observaciones con distribución normal [2]. El ejemplo actual de 30 mediciones de glicemias puede ser probado en el sitio web del programa "graphpad", que se muestra a continuación.

*https://graphpad.com/quickcalcs/Grubbs1.cfm*

Se puede seleccionar la opción de error alfa de 0,05, y luego en la celda copiar los 30 valores hacia abajo, en una columna única y luego solicitar el cálculo. El resultado es el siguiente:

| Row | Value | Z | Significant outlier? |
|---|---|---|---|
| 1 | 56. | 1.53 | |
| 2 | 65. | 1.26 | |
| 3 | 68. | 1.17 | |
| 4 | 75. | 0.96 | |
| 5 | 75. | 0.96 | |
| 6 | 78. | 0.87 | |
| 7 | 78. | 0.87 | |
| 8 | 87. | 0.60 | |
| 9 | 87. | 0.60 | |
| 10 | 87. | 0.60 | |
| 11 | 89. | 0.54 | |
| 12 | 94. | 0.39 | |
| 13 | 99. | 0.24 | |
| 14 | 99. | 0.24 | |
| 15 | 101. | 0.18 | |
| 16 | 104. | 0.09 | |
| 17 | 105. | 0.06 | |
| 18 | 108. | 0.03 | |
| 19 | 115. | 0.24 | |
| 20 | 116. | 0.27 | |
| 21 | 121. | 0.42 | |
| 22 | 122. | 0.45 | |
| 23 | 122. | 0.45 | |
| 24 | 123. | 0.48 | |
| 25 | 132. | 0.75 | |
| 26 | 134. | 0.81 | |
| 27 | 134. | 0.81 | |
| 28 | 140. | 0.99 | |
| 29 | 188. | 2.43 | |
| 30 | 207. | 3.00 | Significant outlier. P < 0.05 |

Se concluye que la medición de 207 es un outlier estadísticamente significativo, que justifica la eliminación de la serie. Se realiza bajo la premisa de que haya una distribución normal.

## *Referencias*


1. *Renze J. Outlier. From MathWorld--A Wolfram Web Resource, created by Eric W. Weisstein. http://mathworld.wolfram.com/Outlier.html*
2. *Grubbs F E. Sample criteria for testing outlying observations. Annals of Mathematical Statistics. 1950; 21 (1): 27–58.*