1. Un outlier se define como una observación que cae más allá de las barras del boxplot, las barras o whiskers se definen como $F_U + 1.5 d_F$ y $F_L - 1.5 d_F$ donde F_U y F_L corresponden a los cuartiles 3 y 1 respectivamente y d_F el rango intercuartil. El extremo superior es el máximo del conjunto de datos

a) ¿Es el extremo superior siempre un outlier?

No, es posible que el extremo superior se encuentre por debajo del whisker $F_U + 1.5 d_F$

b) ¿Es posible para la media o mediana quedar afuera de los cuartiles o incluso afuera de los whiskers?

Por definición, la mediana representa el valor donde se encuentran la mitad de los datos y por lo tanto corresponde al cuartil 2. Como la mediana corresponde al cuartil 2, esta siempre se encontrará a la mitad del rango intercuartil (d_F).

$\therefore$ La mediana no puede quedar afuera de los cuartiles

En cuanto a la media, esta se define como la suma de todos los datos dividida entre los sumandos. Si los datos difieren mucho entre sí, sí puede quedar la media afuera de los cuartiles o incluso de los outliers.

Sea el conjunto de datos

$1, 1, 2, 2, 3, 4, 4, 10^6$ $\qquad n = 8$

$$q_1 = \frac{8+1}{4} = \frac{9}{4} = 2.25 = x_2 + 0.25(x_3 - x_2) = 1 + 0.25(2-1) = 1.25$$

$$\Big\} \; d_F = 2.75$$

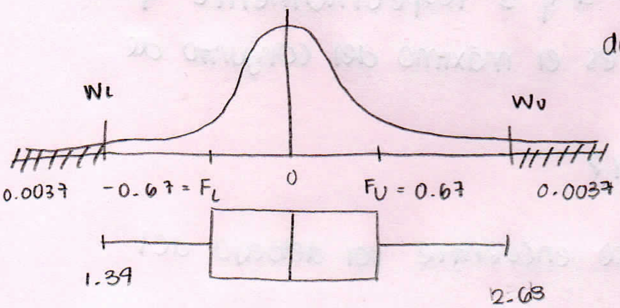
$$q_3 = \frac{3(9)}{4} = \frac{27}{4} = 6.75 = x_6 + 0.75(x_7 * x_6) = 4 + 0.75(4-4) = 4$$

$$\therefore \quad W_U = F_U + 1.5 d_F = 4 + 1.5(2.75) = 8.125$$

$$W_L = F_L - 1.5 d_F = 1.25 - 1.5(2.75) = -2.875$$

Pero $\bar{x} = 125002.125$

c) Sup. que los datos se distribuyen $N(0,1)$, ¿qué porcentaje de los datos se esperan que caigan afuera de los whiskers?



donde F_U: $\mathbb{P}(Z \leq z) = 0.75 \rightarrow z = 0.67$

F_L: $\mathbb{P}(Z \leq z) = 0.25 \rightarrow z = -0.67$

q_2: $\mathbb{P}(Z \leq z) = 0.5 \rightarrow z = 0$

$dF = 2(0.67) = 1.34$

$1.5dF = 2.01$

para los whiskers

$W_U = F_U + 1.5dF = 2.68 \rightarrow \mathbb{P}(Z \leq 2.68) = 0.9963$

$W_L = F_L - 1.5dF = 1.34 \rightarrow \mathbb{P}(////) = 1 - 0.9963 = 0.0037$

$\therefore //// + ///// = 2(0.0037) = 0.0074$

$\therefore$ se espera que el 0.74% de los datos caigan afuera de los whiskers //

d) ¿qué porcentaje de los datos se espera que caigan afuera de los whiskers si suporemos que se distribuyen $N(0,\sigma^2)$ con σ^2 desconocida?

sea $X \sim N(0,1)$

y $W \sim N(0,\sigma^2)$

para F_U: $\mathbb{P}(W \leq w) = \mathbb{P}(\frac{W}{\sigma} \leq \frac{w}{\sigma}) = \mathbb{P}(X \leq x) = 0.75 \rightarrow x = \frac{w}{\sigma} = 0.67 \leftrightarrow w_3 = 0.67\sigma$

F_L: $\mathbb{P}(W \leq w) = \mathbb{P}(\frac{W}{\sigma} \leq \frac{w}{\sigma}) = \mathbb{P}(X \leq x) = 0.25 \rightarrow x = \frac{w}{\sigma} = -0.67 \leftrightarrow w_1 = -0.67\sigma$

De esta manera

$dF = 1.34\sigma$ y $1.5dF = 2.01\sigma$

para los whiskers

W_L: $\mathbb{P}(W \leq 0.67\sigma + 2.01\sigma) = \mathbb{P}(W \leq 2.68\sigma) = \mathbb{P}(\frac{W}{\sigma} \leq 2.68) = 0.9963 \rightarrow 1 - 0.9963 = 0.0037$

$\therefore W_U + W_L = 2(0.0037) = 0.0074$

$\therefore$ se espera que el 0.74% de los datos caigan afuera de los whiskers //