Datos Anómalos

Paúl Arévalo

2024-07-04

Datos Atípicos

Ejemplo

```
datos_bivariados <- read.csv("DatosBivariados.csv")
mean_pesos <- mean(datos_bivariados$Peso)
sd_pesos <- sd(datos_bivariados$Peso)
mean_alturas <- mean(datos_bivariados$Altura)
sd_alturas <- sd(datos_bivariados$Altura)
```

A continuación se genera un conjunto de datos con valores atípicos

```
# set.seed(12345)
# Generar distribución normal con media 50 y desviación estándar 10
pesos <- rnorm(10000, mean = mean_pesos, sd = sd_pesos)
alturas <- rnorm(10000, mean = mean_alturas, sd = sd_alturas)
# Introducir valores atípicos
for (i in 1:100) {
  index <- sample.int(length(pesos), 1)
  pesos[index] <- pesos[index] + round(runif(
    n = 1,
    min = -min(pesos[index] - sd_pesos, sd_pesos + mean_pesos / 2),
    max = sd_pesos + mean_pesos / 2
  ), 0)
}
for (i in 1:100) {
  index <- sample.int(length(alturas), 1)
  alturas[index] <- alturas[index] + round(runif(
    n = 1,
    min = -min(alturas[index] - sd_alturas, sd_alturas + mean_alturas / 2),
    max = sd_alturas + mean_alturas / 2
  ), 0)
}
```

Realizamos una análisis exploratorio de datos:

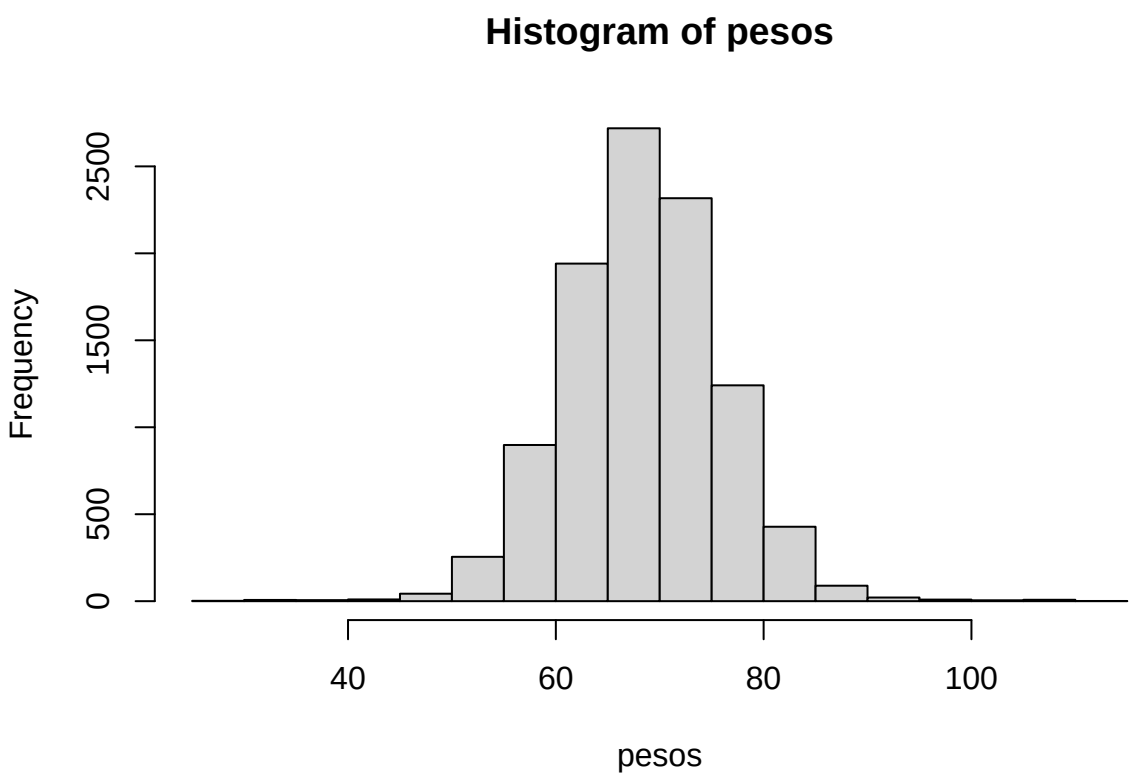
Pesos:

```
# Observar la distribución de los datos
summary(pesos)
```

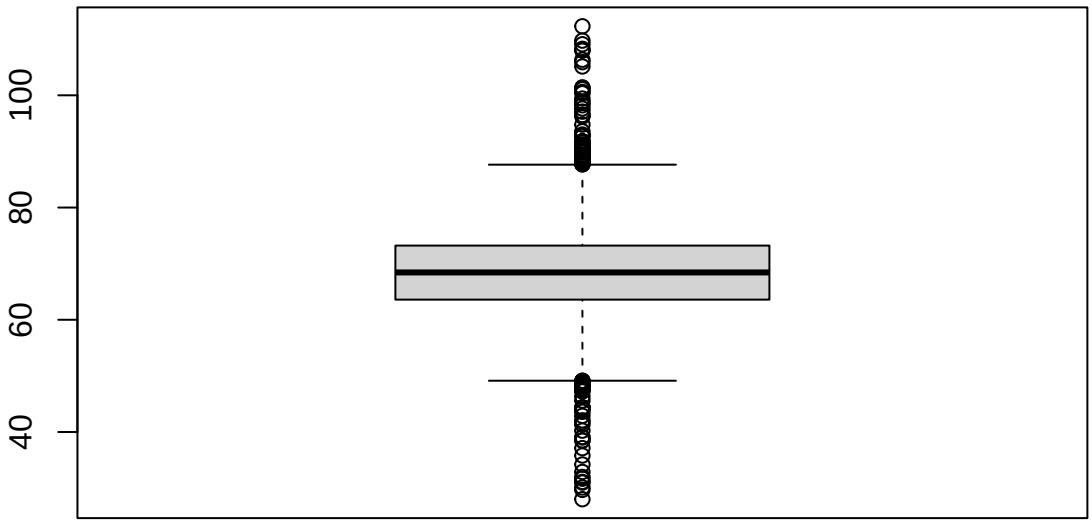
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   28.03   63.59   68.43   68.44   73.22  112.30
```

```
hist(pesos)
```




```
boxplot(pesos)
```

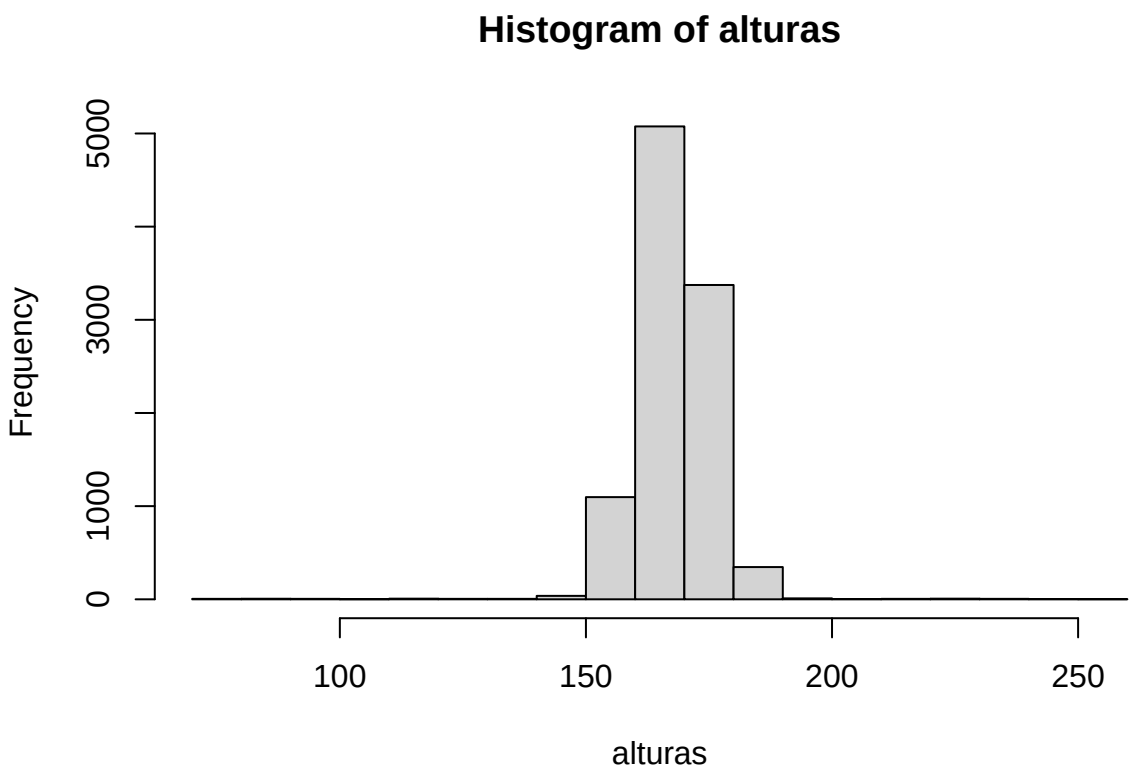


Alturas

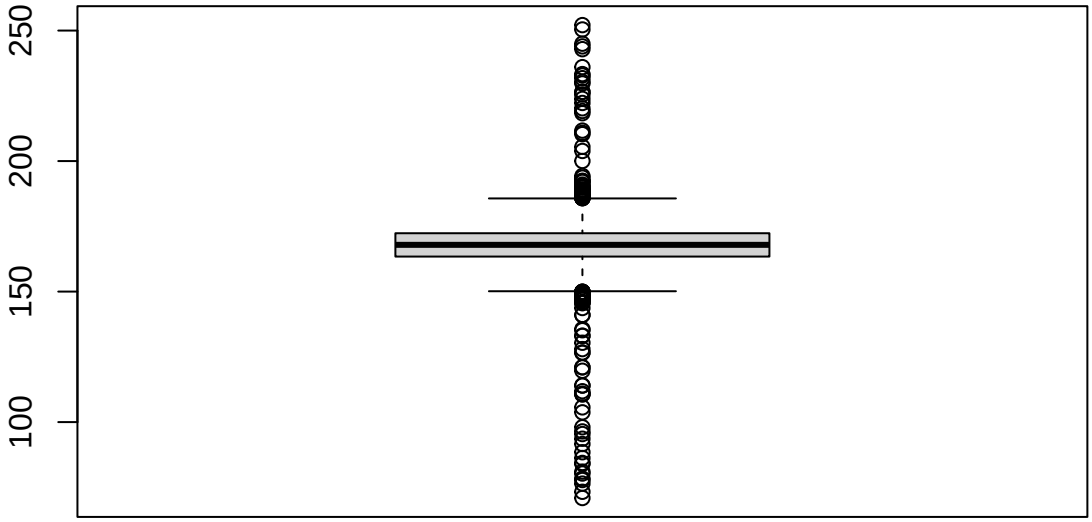
```
# Observar la distribución de los datos
summary(alturas)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   70.92  163.42  167.92  167.84  172.36  252.10
```

```r
hist(alturas)
```




```r
boxplot(alturas)
```

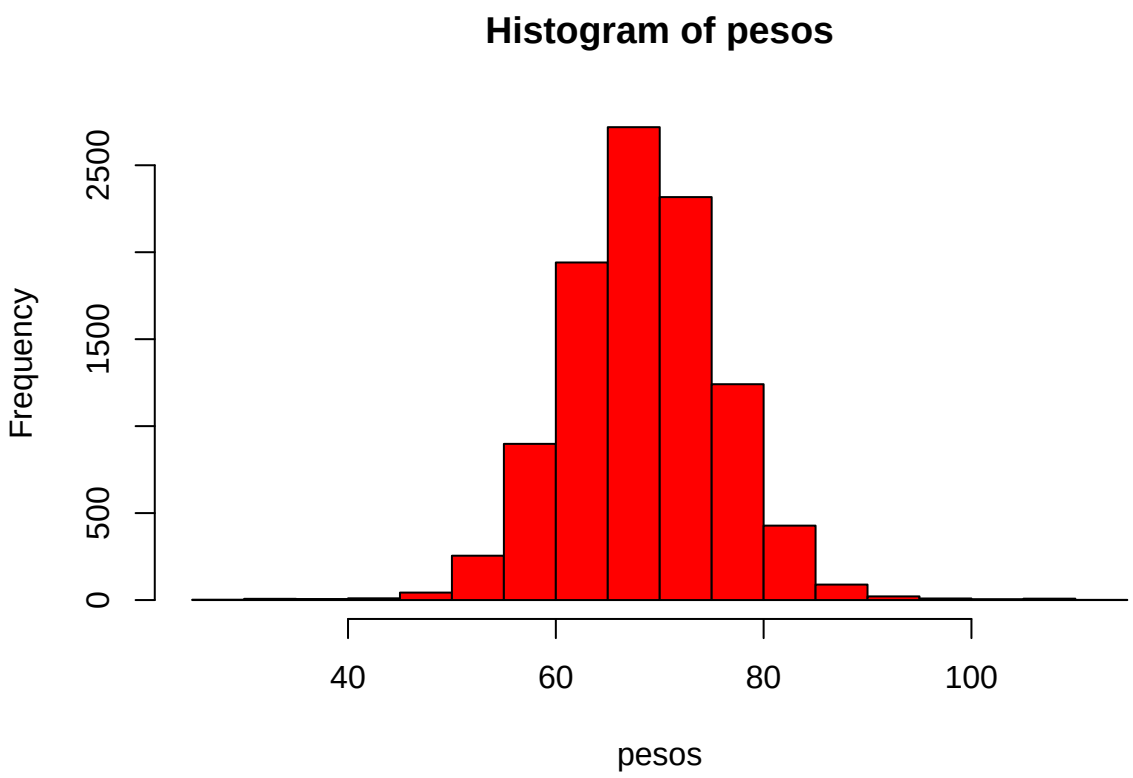


Procedemos a detectar valores atípicos

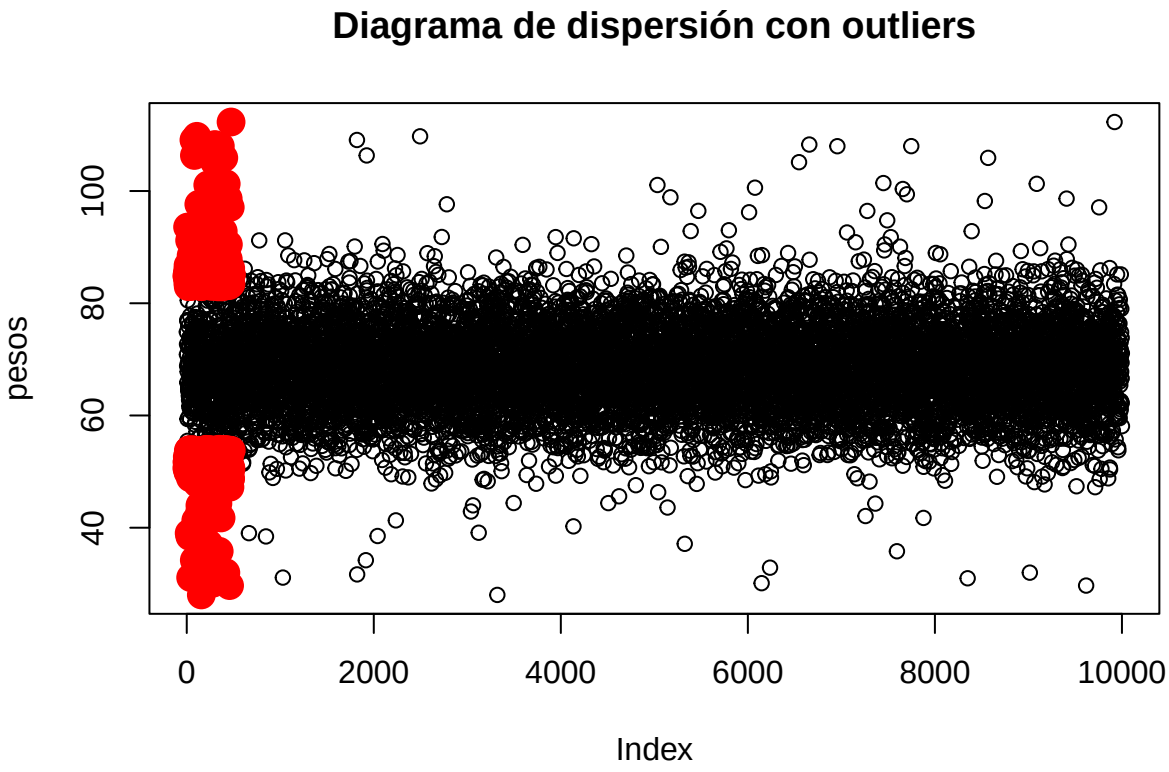
Pesos

```r
# Utilizar el método IQR (Rango intercuartílico)
iqr <- IQR(pesos)
limite_superior <- median(pesos) + 1.5 * iqr
limite_inferior <- median(pesos) - 1.5 * iqr
```

```r
outliers <- pesos[pesos > limite_superior | pesos < limite_inferior]
# Identificar outliers en el histograma
hist(pesos, col = ifelse(is.na(outliers), "lightblue", "red"))
```

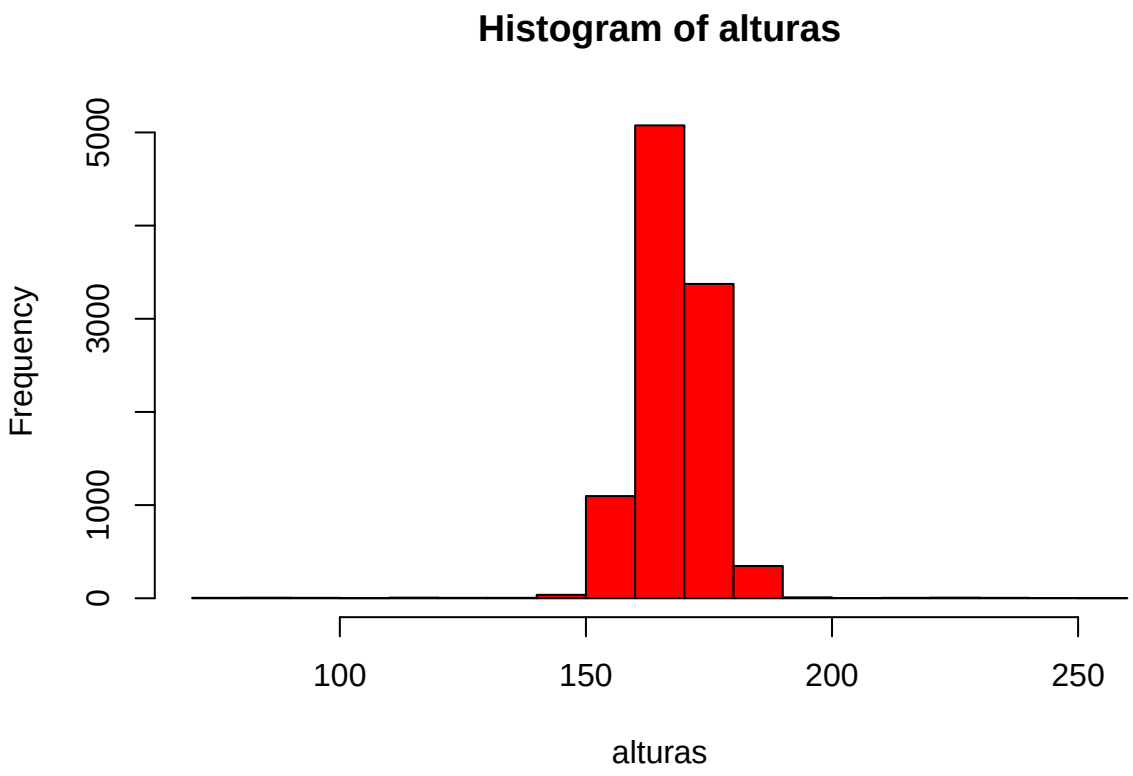



```r
plot(pesos, main = "Diagrama de dispersión con outliers")
points(outliers, col = "red", pch = 16, cex = 2)
```



Alturas

```
# Utilizar el método IQR (Rango intercuartílico)
iqr <- IQR(alturas)
limite_superior <- median(alturas) + 1.5 * iqr
limite_inferior <- median(alturas) - 1.5 * iqr
outliers <- alturas[alturas > limite_superior | alturas < limite_inferior]
# Identificar outliers en el histograma
hist(alturas, col = ifelse(is.na(outliers), "lightblue", "red"))
```




```r
plot(alturas, main = "Diagrama de dispersión con outliers")
points(outliers, col = "red", pch = 16, cex = 2)
```