AUA CS108, Statistics, Fall 2020

Lecture 11

Michael Poghosyan

18 Sep 2020

Contents

BoxPlot, Example

Hovhannes's Problem: Assume 50% of our data is 0, 25% is -1 and 25% is 1. Are all -1's and 1's Outliers?

BoxPlot, Example

Question: Last time there was a question about how **R** is calculating `summary`.

BoxPlot, Example

Question: Last time there was a question about how **R** is calculating `summary`. Well, although I do not know yet how it is calculated, I can show what is doing `fivenum` command:

BoxPlot, Example

Question: Last time there was a question about how **R** is calculating `summary`. Well, although I do not know yet how it is calculated, I can show what is doing `fivenum` command:

```
fivenum
```

```
## function (x, na.rm = TRUE)
## {
##     xna <- is.na(x)
##     if (any(xna)) {
##         if (na.rm)
##             x <- x[!xna]
##         else return(rep.int(NA, 5))
##     }
##     x <- sort(x)
##     n <- length(x)
##     if (n == 0)
##         rep.int(NA, 5)
##     else {
##         n4 <- floor((n + 3)/2)/2
##         d <- c(1, n4, (n + 1)/2, n + 1 - n4, n)
##         0.5 * (x[floor(d)] + x[ceiling(d)])
##     }
## }
## <bytecode: 0x0000000007eeac68>
## <environment: namespace:stats>
```

BoxPlot, Common Errors

Here is a common error when Plotting the BoxPlot:

BoxPlot, Common Errors

Here is a common error when Plotting the BoxPlot:

- One uses $W_1 = Q_1 - 1.5 \cdot IQR$ and $W_2 = Q_3 + 1.5 \cdot IQR$. This is **not correct**!

BoxPlot, Common Errors

Here is a common error when Plotting the BoxPlot:

- One uses $W_1 = Q_1 - 1.5 \cdot IQR$ and $W_2 = Q_3 + 1.5 \cdot IQR$. This is **not correct**! W_1 and W_2 need to be from our Dataset!

BoxPlot, Common Errors

Here is a common error when Plotting the BoxPlot:

- One uses $W_1 = Q_1 - 1.5 \cdot IQR$ and $W_2 = Q_3 + 1.5 \cdot IQR$. This is **not correct**! W_1 and W_2 need to be from our Dataset!

Take as W_1 and W_2 the smallest and largest **Datapoints**, respectively, in

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right].$$

BoxPlot, Common Errors

Here is a common error when Plotting the BoxPlot:

- One uses $W_1 = Q_1 - 1.5 \cdot IQR$ and $W_2 = Q_3 + 1.5 \cdot IQR$. This is **not correct**! W_1 and W_2 need to be from our Dataset!

Take as W_1 and W_2 the smallest and largest **Datapoints**, respectively, in

$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right].$$

And an important

Note: always keep the scale on the x-axis! Place the numbers in correct places, keep the distance between numbers.

Additions/Variations:

Some Variations:

- Variable Width BoxPlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot
- VasePlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot
- VasePlot
- ViolinPlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot
- VasePlot
- ViolinPlot
- BeanPlot

Additions/Variations:

Some Variations:

- Variable Width BoxPlot
- Notched BoxPlot
- VasePlot
- ViolinPlot
- BeanPlot

See, for Example, this page.

Boxplot, Why we use it

We use BoxPlots to:

Boxplot, Why we use it

We use BoxPlots to:

- Visualize the distribution of the Dataset

Boxplot, Why we use it

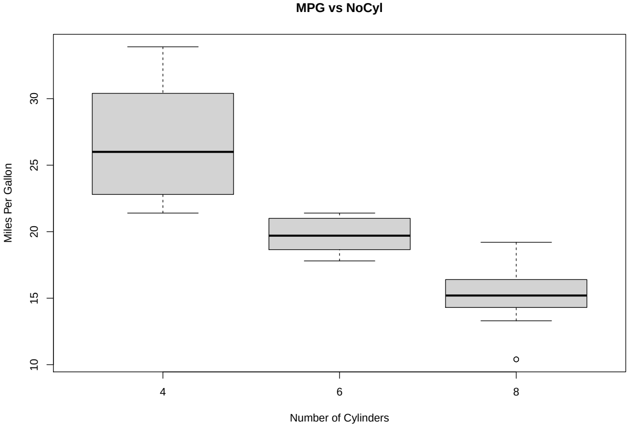
We use BoxPlots to:

- Visualize the distribution of the Dataset
- To compare two or more Datasets

Example

Here we use the `mtcars` Dataset:

```
boxplot( mpg~cyl, data=mtcars, main="MPG vs NoCyl",
   xlab="Number of Cylinders", ylab="Miles Per Gallon")
```

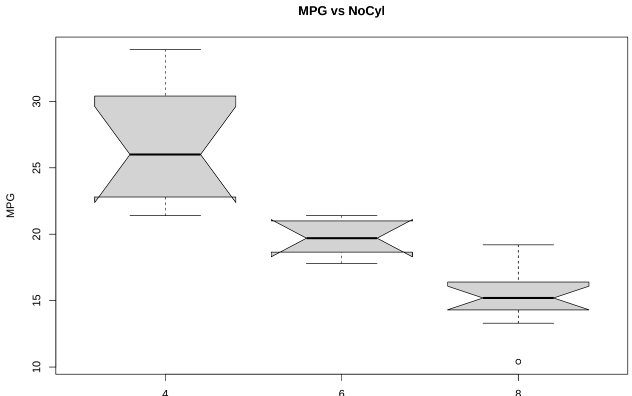


Example

Again,

```
boxplot( mpg~cyl, data=mtcars, notch = T,
         main="MPG vs NoCyl", xlab="Number of Cylinders", y
```

```
## Warning in bxp(list(stats = structure(c(21.4, 22.8, 26,
## notches went outside hinges ('box'): maybe set notch=FAI
```



Example

Few days ago we all received AUA Insider | News email.

Example

Few days ago we all received AUA Insider | News email. The following is from that email: Link

Example

Few days ago we all received AUA Insider | News email. The following is from that email: Link

This is an Exploratory Analysis for the Kangaroo, Meghu and Russian Bear Cub contests results in Armenia and Artsakh. The Shiny app (created by **R**), is here: link.

BoxPlot, Notes

Note: Recall that an **Outlier** in the BoxPlot sense is a Datapoint x_k with

$$x_k \notin \left[Q_1 - \frac{3}{2} IQR,\ Q_3 + \frac{3}{2} IQR \right].$$

BoxPlot, Notes

Note: Recall that an **Outlier** in the BoxPlot sense is a Datapoint x_k with

$$x_k \notin \left[Q_1 - \frac{3}{2} IQR,\ Q_3 + \frac{3}{2} IQR \right].$$

Another way to define an **Outlier:** Datapoint x_k is an Outlier, if

$$|x_k - \bar{x}| \geq 3 \cdot sd(x).$$

BoxPlot, Notes

Note: Recall that an **Outlier** in the BoxPlot sense is a Datapoint x_k with

$$x_k \notin \left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right].$$

Another way to define an **Outlier:** Datapoint x_k is an Outlier, if

$$|x_k - \bar{x}| \geq 3 \cdot sd(x).$$

Note: Where the coefficient $\dfrac{3}{2}$ in front of the IQR comes from?

BoxPlot, Notes

Note: Recall that an **Outlier** in the BoxPlot sense is a Datapoint x_k with

$$x_k \notin \left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right].$$

Another way to define an **Outlier:** Datapoint x_k is an Outlier, if

$$|x_k - \bar{x}| \geq 3 \cdot sd(x).$$

Note: Where the coefficient $\dfrac{3}{2}$ in front of the IQR comes from? This comes from the Normal Distribution: if our r.v. X is Normally Distributed, then (with theoretical Quartiles)

$$\mathbb{P}(X \in [Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]) \approx 0.993,$$

so the chances that an Observation will be outside of this interval are very small.

BoxPlot, Notes

Note: Recall that an **Outlier** in the BoxPlot sense is a Datapoint x_k with

$$x_k \notin \left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right].$$

Another way to define an **Outlier:** Datapoint x_k is an Outlier, if

$$|x_k - \bar{x}| \geq 3 \cdot sd(x).$$

Note: Where the coefficient $\dfrac{3}{2}$ in front of the IQR comes from? This comes from the Normal Distribution: if our r.v. X is Normally Distributed, then (with theoretical Quartiles)

$$\mathbb{P}(X \in [Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]) \approx 0.993,$$

so the chances that an Observation will be outside of this interval are very small. So if we see that kind of Observation, we think that this number is an Outlier.

BoxPlot, Notes

Note: Sometimes, BoxPlot's Whiskers span to the Max and Min Datapoints, so in this case BoxPlot doesn't show Outliers.