AUA CS 108, Statistics, Fall 2019

Lecture 09

Michael Poghosyan
YSU, AUA
michael@ysu.am, mpoghosyan@aua.am

Friday the 13th, Sep 2019

Contents

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation
- Give the Definition of the MAD

Last Lecture ReCap

- Define the Sample Variance and the Standard Deviation
- Give the Definition of the MAD
- What are the Quartiles?

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation.

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences
  $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and
  $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,i.e., the first and last observations lying in
$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right];$$

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,i.e., the first and last observations lying in
$$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right];$$

the line joining that fences to corresponding quartiles are the *Whiskers*;

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualisation. To draw the BoxPlot, we calculate the following:

- The Quartiles $Q_1, Q_2 = Median, Q_3$
- the Lower and Upper Fences
  $W_1 = \min\{x_i : x_i \geq Q_1 - 1.5 \cdot IQR\}$ and
  $W_2 = \max\{x_i : x_i \leq Q_3 + 1.5 \cdot IQR\}$,i.e., the first and last observations lying in
  $$\left[Q_1 - \frac{3}{2}IQR,\ Q_3 + \frac{3}{2}IQR\right];$$

the line joining that fences to corresponding quartiles are the *Whiskers*;

- the set of all Outliers
  $$O = \left\{x_i : x_i \notin \left[Q_1 - \frac{3}{2}IQR, Q_3 + \frac{3}{2}IQR\right]\right\}$$

BoxPlot, Example

Then we draw the points W_1, Q_1, Q_2, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$.

BoxPlot, Example

Then we draw the points W_1, Q_1, Q_2, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$.

Example: Draw the Boxplot of

$$x: \; 0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12$$

BoxPlot, Example

Then we draw the points W_1, Q_1, Q_2, Q_3, W_2 on the real line and add all outliers, and make a box over $[Q_1, Q_3]$.

Example: Draw the Boxplot of

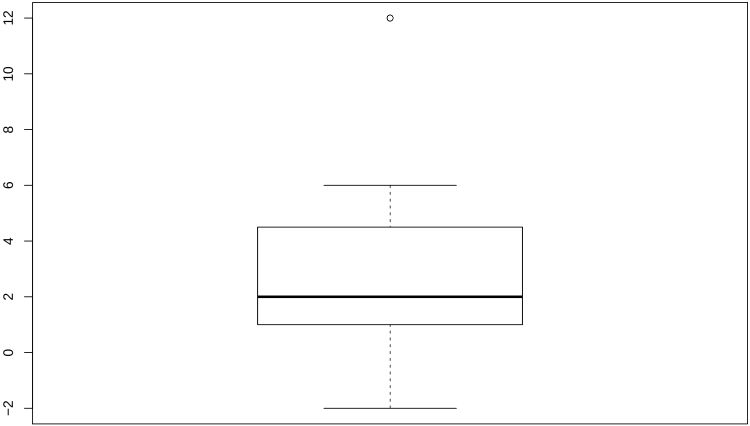
$$x: \; 0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12$$

Solution: OTB;

BoxPlot, Example

Now, using **R**:

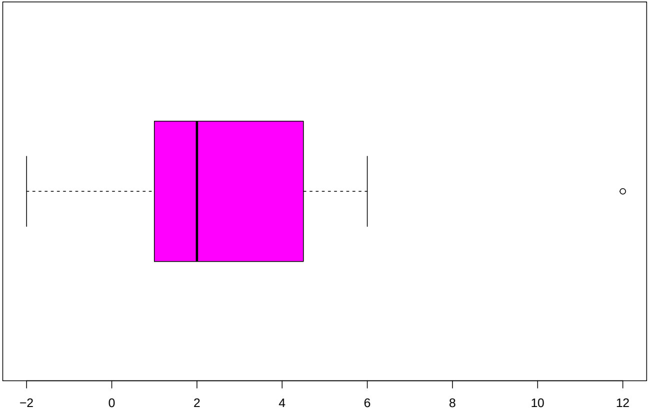
```r
x <- c(0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12)
boxplot(x)
```



BoxPlot, Example

Another view:

```
x <- c(0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12)
boxplot(x, horizontal = T, col = "magenta")
```



BoxPlot, Common Errors

Here are some common errors when Plotting the BoxPlot:

BoxPlot, Common Errors

Here are some common errors 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 are some common errors 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 are some common errors 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! This is because by the Box we want to show the Spread of the middle-half of our Dataset.

BoxPlot, Common Errors

Here are some common errors 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! This is because by the Box we want to show the Spread of the middle-half of 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 are some common errors 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! This is because by the Box we want to show the Spread of the middle-half of our Dataset.

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

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

- One doesn't keep the scale.

BoxPlot, Common Errors

Here are some common errors 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! This is because by the Box we want to show the Spread of the middle-half of our Dataset.

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

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

- One doesn't keep the scale. Say, one draws the Median exactly at the middle of the Quartiles, despite the Dataset is not symmetric at all.

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

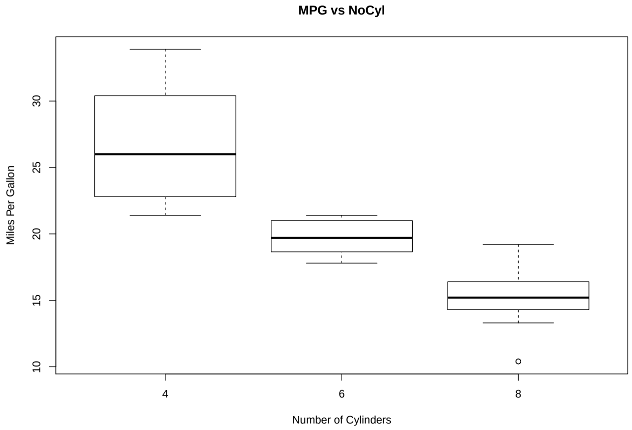
We use BoxPlots to:

- 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")
```

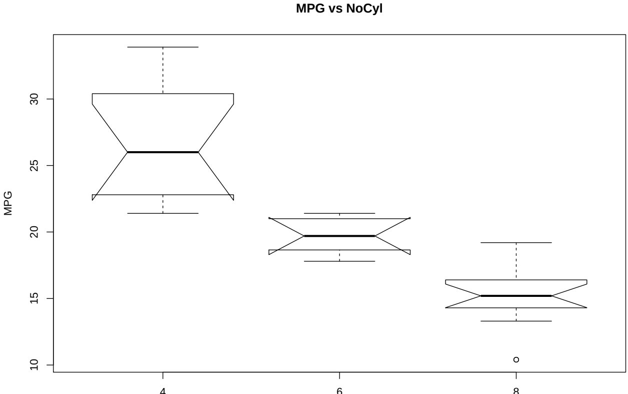


Example

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
```



Note

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

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

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

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

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles.

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Now, let $\alpha \in (0, 1)$.

Sample Quantiles

Now we want to generalize the idea of the Median and Quartiles. Recall that:

- 50% of Datapoints are to the left of the Median, and 50% are to the right, so Median divides the (sorted) Dataset in the (approximate) proportion 50% - 50%
- 25% of Datapoints are to the left of the Lower Quartile Q_1, and 75% are to the right, so Q_1 divides the (sorted) Dataset in the (approximate) proportion 25%-75%
- 75% of Datapoints are to the left of the Upper Quartile Q_3, and 25% are to the right, so Q_3 divides the (sorted) Dataset in the (approximate) proportion 75%-25%

Now, let $\alpha \in (0, 1)$. We want to find a real number q_α dividing our (sorted) Dataset into the proportion $100\alpha\% - 100(1 - \alpha)\%$, i.e., q_α is a point such that the α-portion of the Datapoints are to the left to q_α, and others are to the right.

Sample Quantiles

Let $x: \; x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0, 1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Sample Quantiles

Let $x: \ x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0, 1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: $[\alpha \cdot n]$ is the integer part of $\alpha \cdot n$, and $x_{([\alpha \cdot n])}$ is the $[\alpha \cdot n]$-th Order Statistics.

Sample Quantiles

Let $x: \ x_1, x_2, ..., x_n$ be our 1D numerical Dataset. Assume also that $\alpha \in (0,1)$.

Definition: The Quantile of order α (or $100\alpha\%$ order, the α-Quantile) of x is defined by

$$q_\alpha = q_\alpha^x = x_{([\alpha \cdot n])}.$$

Note: $[\alpha \cdot n]$ is the integer part of $\alpha \cdot n$, and $x_{([\alpha \cdot n])}$ is the $[\alpha \cdot n]$-th Order Statistics.

Note: There are different definitions of the α-quantile in the literature and in software implementations. Say, **R** has 9 methods to calculate quantiles.