ASDS Statistics, YSU, Fall 2020

Lecture 09

Michael Poghosyan

03 Oct 2020

Contents

Last Lecture Recap

- Name some Statistics for the Spread/Variability of a Dataset

Last Lecture Recap

- Name some Statistics for the Spread/Variability of a Dataset
- Define the Deviations and Absolute Deviations from the Mean

Last Lecture Recap

- Name some Statistics for the Spread/Variability of a Dataset
- Define the Deviations and Absolute Deviations from the Mean
- Give the Definition of the Sample Variance;

Last Lecture Recap

- Name some Statistics for the Spread/Variability of a Dataset
- Define the Deviations and Absolute Deviations from the Mean
- Give the Definition of the Sample Variance;
- Give the Definition of the Sample SD;

Last Lecture Recap

- Name some Statistics for the Spread/Variability of a Dataset
- Define the Deviations and Absolute Deviations from the Mean
- Give the Definition of the Sample Variance;
- Give the Definition of the Sample SD;
- Give the Definition of the MAD;

Last Lecture Recap

- Name some Statistics for the Spread/Variability of a Dataset
- Define the Deviations and Absolute Deviations from the Mean
- Give the Definition of the Sample Variance;
- Give the Definition of the Sample SD;
- Give the Definition of the MAD;
- What is the idea behind the Quartiles?

Last Lecture Recap

- Name some Statistics for the Spread/Variability of a Dataset
- Define the Deviations and Absolute Deviations from the Mean
- Give the Definition of the Sample Variance;
- Give the Definition of the Sample SD;
- Give the Definition of the MAD;
- What is the idea behind the Quartiles?
- Define the IQR.

Quartiles and IQR

Remark: Note that the Quartiles Q_1, Q_2, Q_3 are not always Datapoints.

Quartiles and IQR

Remark: Note that the Quartiles Q_1, Q_2, Q_3 are not always Datapoints.

Note: Recall the idea of Quartiles: the points Q_1, Q_2, Q_3 on the real axis divide our Dataset into (almost) four equal-length portions:

- almost 25% of our Datapoints are to the left to Q_1

Quartiles and IQR

Remark: Note that the Quartiles Q_1, Q_2, Q_3 are not always Datapoints.

Note: Recall the idea of Quartiles: the points Q_1, Q_2, Q_3 on the real axis divide our Dataset into (almost) four equal-length portions:

- almost 25% of our Datapoints are to the left to Q_1
- almost 25% of our Datapoints are between Q_1 and Q_2

Quartiles and IQR

Remark: Note that the Quartiles Q_1, Q_2, Q_3 are not always Datapoints.

Note: Recall the idea of Quartiles: the points Q_1, Q_2, Q_3 on the real axis divide our Dataset into (almost) four equal-length portions:

- almost 25% of our Datapoints are to the left to Q_1
- almost 25% of our Datapoints are between Q_1 and Q_2
- almost 25% of our Datapoints are between Q_2 and Q_3

Quartiles and IQR

Remark: Note that the Quartiles Q_1, Q_2, Q_3 are not always Datapoints.

Note: Recall the idea of Quartiles: the points Q_1, Q_2, Q_3 on the real axis divide our Dataset into (almost) four equal-length portions:

- almost 25% of our Datapoints are to the left to Q_1
- almost 25% of our Datapoints are between Q_1 and Q_2
- almost 25% of our Datapoints are between Q_2 and Q_3
- almost 25% of our Datapoints are to the right to Q_3

Quartiles and IQR

Remark: Note that the Quartiles Q_1, Q_2, Q_3 are not always Datapoints.

Note: Recall the idea of Quartiles: the points Q_1, Q_2, Q_3 on the real axis divide our Dataset into (almost) four equal-length portions:

- almost 25% of our Datapoints are to the left to Q_1
- almost 25% of our Datapoints are between Q_1 and Q_2
- almost 25% of our Datapoints are between Q_2 and Q_3
- almost 25% of our Datapoints are to the right to Q_3

Note: The interval $[Q_1, Q_3]$ contains almost the half of the Datapoints.

Quartiles and IQR

Remark: Note that the Quartiles Q_1, Q_2, Q_3 are not always Datapoints.

Note: Recall the idea of Quartiles: the points Q_1, Q_2, Q_3 on the real axis divide our Dataset into (almost) four equal-length portions:

- almost 25% of our Datapoints are to the left to Q_1
- almost 25% of our Datapoints are between Q_1 and Q_2
- almost 25% of our Datapoints are between Q_2 and Q_3
- almost 25% of our Datapoints are to the right to Q_3

Note: The interval $[Q_1, Q_3]$ contains almost the half of the Datapoints. So the IQR shows the Spread of the middle half of our Dataset, it is a measure of the Spread/Variability.

Quartiles in **R**

In **R**, one can use the commands `quantile(x, 0.25)` and `quantile(x, 0.75)` to find Q_1 and Q_3.

Quartiles in **R**

In **R**, one can use the commands `quantile(x, 0.25)` and `quantile(x, 0.75)` to find Q_1 and Q_3. For example,

```r
x <- 1:10
quantile(x,0.25)
```

```
##  25%
## 3.25
```

Quartiles in R

In **R**, one can use the commands `quantile(x, 0.25)` and `quantile(x, 0.75)` to find Q_1 and Q_3. For example,

```
x <- 1:10
quantile(x,0.25)
```

```
##  25%
## 3.25
```

If you will not give a parameter to `quantile`, **R** will calculate 0% (minimum datapoint), 25%, 50%, 75% and 100% (maximum datapoint) quartiles:

```
x <- 1:10
quantile(x)
```

```
##    0%   25%   50%   75%  100%
##  1.00  3.25  5.50  7.75 10.00
```

Quartiles in R

Also, you can use the following commands:

```
x <- 1:10
fivenum(x)
```

```
## [1]  1.0  3.0  5.5  8.0 10.0
```

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    3.25    5.50    5.50    7.75   10.00
```

Quartiles in **R**

Also, you can use the following commands:

```
x <- 1:10
fivenum(x)
```

```
## [1]  1.0  3.0  5.5  8.0 10.0
```

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.25    5.50    5.50    7.75   10.00
```

To calculate the IQR in **R**, we can use the `IQR` command:

```
x <- 1:10
IQR(x)
```

```
## [1] 4.5
```

Note

Note: Please note that **R** is not using our definition of the Quartiles, so sometimes we will get different results when calculating by a hand or by **R**.

BoxPlot

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

BoxPlot

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

BoxPlot

BoxPlot (or Box and Whiskers Plot) is another very common method of visualization. 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 visualization. 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 visualization. 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 visualization. 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 lines joining that fences to corresponding quartiles are the *Whiskers*;

BoxPlot

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

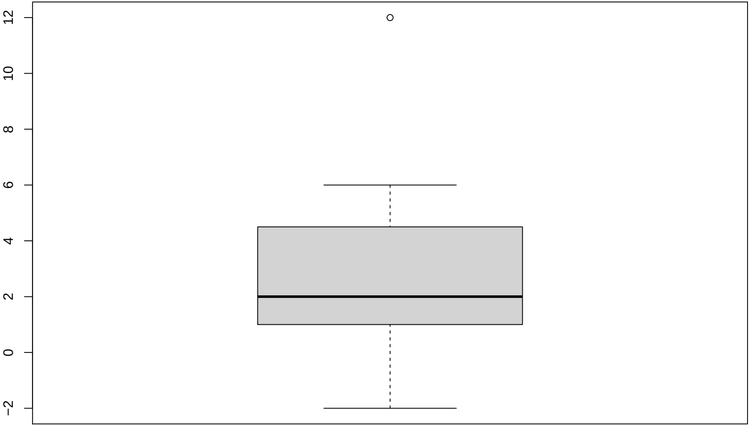
$$x: \; 0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12$$

Solution: OTB

BoxPlot, Example

Now, using **R**:

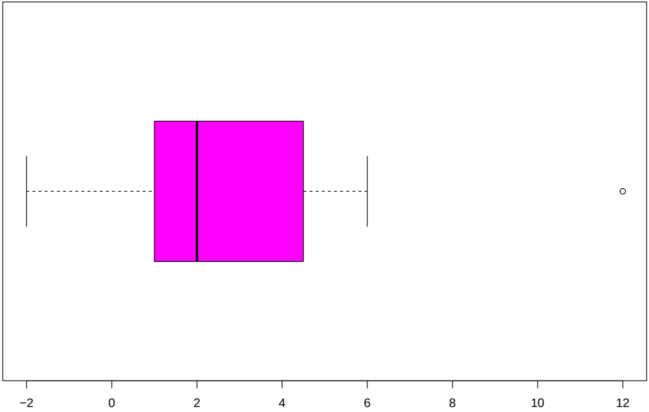
```r
x <- c(0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12)
boxplot(x)
```



BoxPlot, Example

Another view:

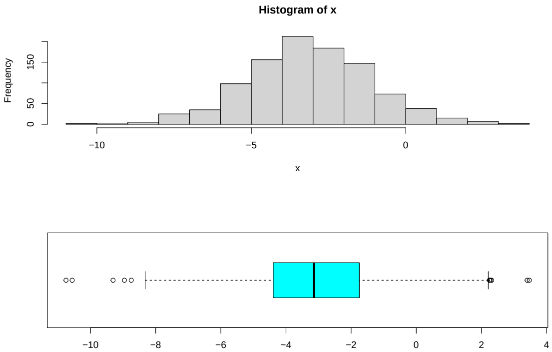
```
x <- c(0, -2, 2, 1, 5, 6, 4, 1, 2, 1, 12)
boxplot(x, horizontal = T, col = "magenta")
```



BoxPlot, Example

Here are some Datasets' Histograms along with the BoxPlots:

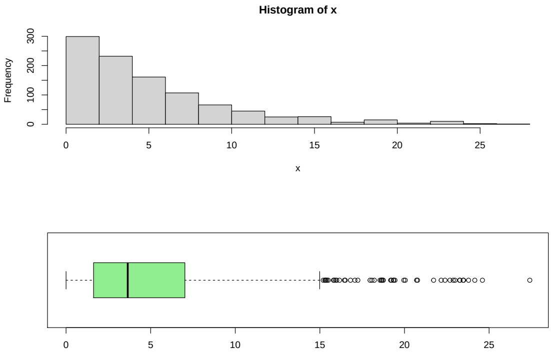
```
x <- rnorm(1000, mean = -3, sd = 2)
par(mfrow=c(2,1)); hist(x)
boxplot(x, horizontal = T, col = "cyan")
```



BoxPlot, Example

Here are some Datasets' Histograms along with the BoxPlots:

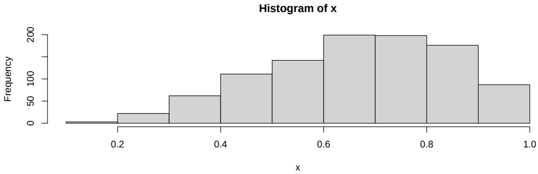
```
x <- rexp(1000, rate = 0.2)
par(mfrow=c(2,1)); hist(x)
boxplot(x, horizontal = T, col = "lightgreen")
```

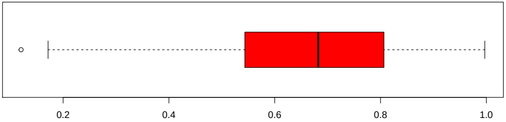


BoxPlot, Example

Here are some Datasets' Histograms along with the BoxPlots:

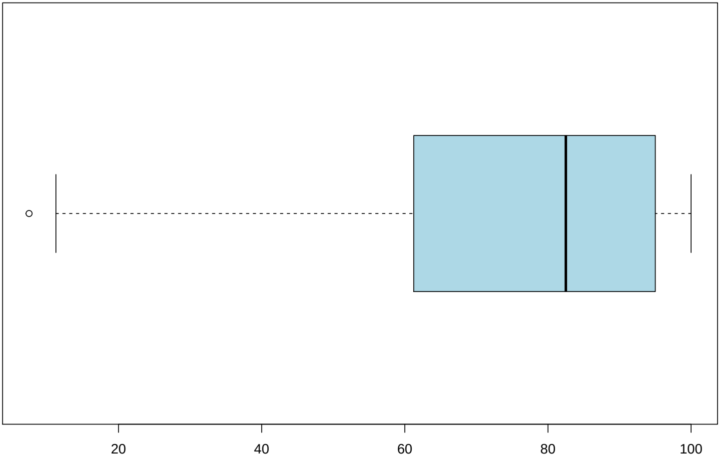
```
x <- rbeta(1000, shape1 = 4, shape2 = 2)
par(mfrow=c(2,1)); hist(x)
boxplot(x, horizontal = T, col = "red")
```





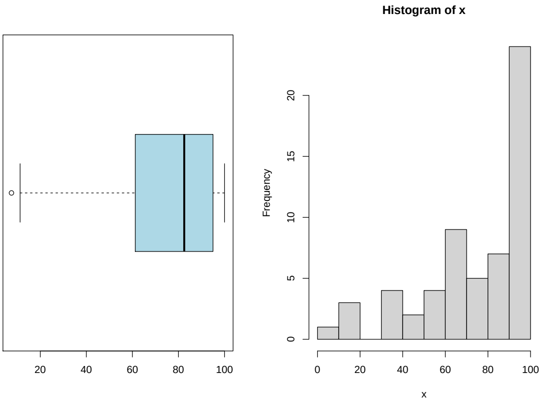
BoxPlot, Example

Here is the BoxPlot of the AUA Stat Quiz grades: can you describe the result?



BoxPlot, Example

And here is the BoxPlot of the same Quiz grades along with the Histogram:




```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.50   61.25   82.50   74.63   95.00  100.00
```

BoxPlot, Common Error

Here is a common error when Plotting the BoxPlot:

BoxPlot, Common Error

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 Error

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 Error

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

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

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

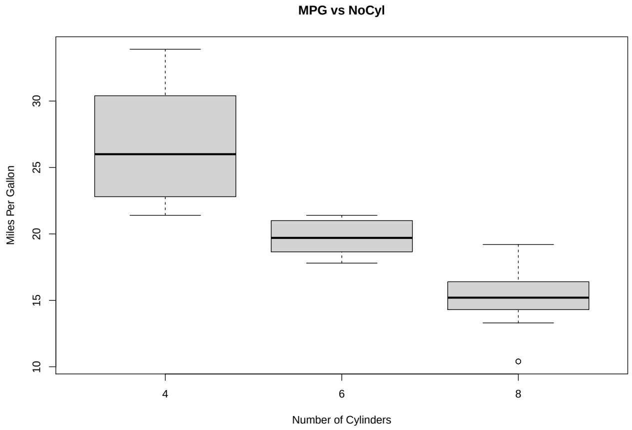
We use BoxPlots to:

- 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")
```

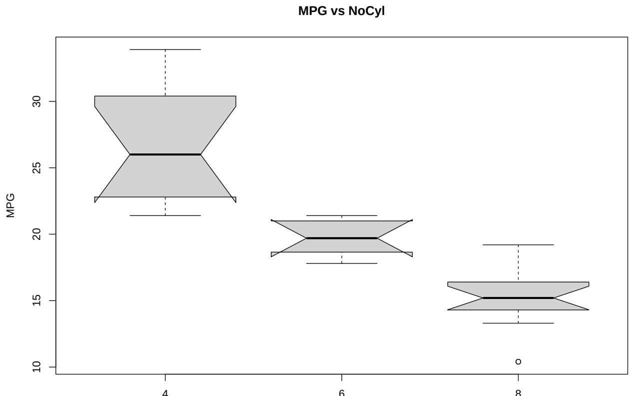


Example

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=FAL
```



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

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

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

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 $\frac{3}{2}$ 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.

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.