AUA CS 108, Statistics, Fall 2019

Lecture 06

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

06 Sep 2019

Contents

- Stem'n'leaf Plot
- Scatter Plot
- Order Statistics

Last Lecture ReCap

- Is it always good to have small (width) bins for a Histogram?

Last Lecture ReCap

- Is it always good to have small (width) bins for a Histogram?
- How to add Datapoints to the Histogram in **R**?

Last Lecture ReCap

- Is it always good to have small (width) bins for a Histogram?
- How to add Datapoints to the Histogram in **R**?
- What is the **KDE**

Last Lecture ReCap

- Is it always good to have small (width) bins for a Histogram?
- How to add Datapoints to the Histogram in **R**?
- What is the **KDE**
- Give the definition of **KDE**.

Last Lecture ReCap

- Is it always good to have small (width) bins for a Histogram?
- How to add Datapoints to the Histogram in **R**?
- What is the **KDE**
- Give the definition of **KDE**.
- How to consruct the **S-n-L Plot**?

Last Lecture ReCap

- Is it always good to have small (width) bins for a Histogram?
- How to add Datapoints to the Histogram in **R**?
- What is the **KDE**
- Give the definition of **KDE**.
- How to consruct the **S-n-L Plot**?
- Why is it for?

Example, SnL Plot

This is from our last lecture: we use again the *airquality* Dataset, but now, the *Wind* Variable:

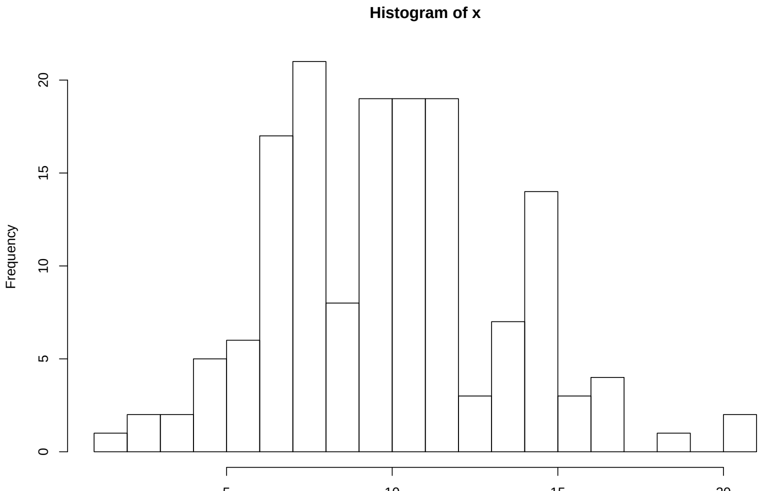
```
x <- airquality$Wind
stem(x)
```

```
##
##   The decimal point is at the |
##
##    1 | 7
##    2 | 38
##    3 | 4
##    4 | 016666
##    5 | 111777
##    6 | 3333333399999999
##    7 | 4444444444
##    8 | 0000000000066666666
##    9 | 2222222277777777777
##   10 | 3333333333399999999
##   11 | 555555555555555
##   12 | 0000666
##   13 | 2288888
##   14 | 333333999999999
##   15 | 555
##   16 | 1666
##   17 |
##   18 | 4
##   19 |
##   20 | 17
```

Example, SnL Plot

Let's draw the Histogram of the same Dataset:

```r
x <- airquality$Wind
hist(x, breaks = 15)
```



Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram

Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram
- Pros of SnL is that

Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram
- Pros of SnL is that we can recover the Dataset from it (if no rounding was made), but not from the Histogram

Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram
- Pros of SnL is that we can recover the Dataset from it (if no rounding was made), but not from the Histogram
- Cons of SnL is that it is for a small-size Dataset

Notes

- From the previous 2 graphs, you can see that SnL Plot is, in some sense, an inverted Histogram
- Pros of SnL is that we can recover the Dataset from it (if no rounding was made), but not from the Histogram
- Cons of SnL is that it is for a small-size Dataset

Say, you can try

```r
x <- rnorm(10000)
stem(x)
```

Some Parameters of the SnL Plot

Let's run the following code:

```
set.seed(77777)
x <- sample(1:30, size = 20, replace = T)
stem(x)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 1113
##   0 | 6689
##   1 | 0023333
##   1 | 9
##   2 | 0124
```

Some Parameters of the SnL Plot

Let's runt the following code:

```
set.seed(77777)
x <- sample(1:30, size = 20, replace = T)
stem(x, scale = 2)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 1113
##   0 | 6689
##   1 | 0023333
##   1 | 9
##   2 | 0124
```

Some Parameters of the SnL Plot

Let's runt the following code:

```r
set.seed(77777)
x <- sample(1:30, size = 20, replace = T)
stem(x, scale = 0.5)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 11136689
##   1 | 00233339
##   2 | 0124
```

Some Additions: Comparing 2 Groups, Back-to-Back Histograms and SnL Plots

Sometimes we want to compare the values of the same variable for two different groups, say, the Height Variable for the Man and Woman groups.

Some Additions: Comparing 2 Groups, Back-to-Back Histograms and SnL Plots

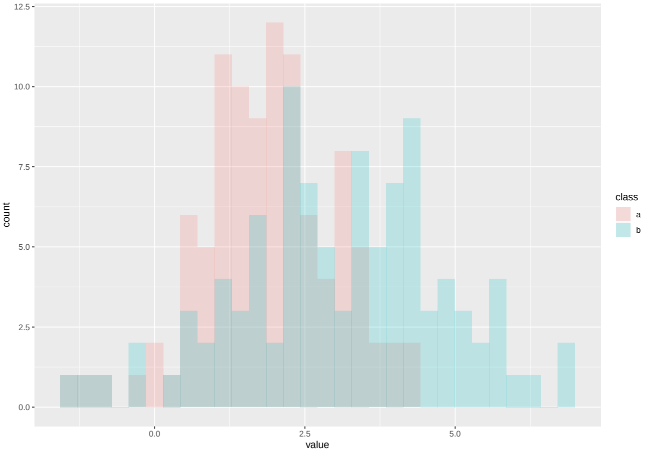
Sometimes we want to compare the values of the same variable for two different groups, say, the Height Variable for the Man and Woman groups. Then, we can use different colors to visualize the difference.

Example

Here is a synthetic (artifical) example:

```
library(ggplot2)
v1 <- rnorm(100,2,1); v2 <- rnorm(100,3,2)
df <- data.frame(value = c(v1, v2), class = rep(c("a","b"), each=100))
ggplot(df, aes(x=value, fill=class)) + geom_histogram(alpha=0.2, position="identity")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Example

And sometimes Back-to-back Histograms, SnL Plots or Barplots can help.

Example

We do not have a command to draw a Back-to-Back SnL Plot, so we load the *aplpack* package:

```
x <- sample(1:30, size = 50, replace = T);
y <- sample(1:30, size = 50, replace = T);
aplpack::stem.leaf.backback(x,y,  rule.line = "Sturges")
```

```
## ________________________________________________
##   1 | 2: represents 12, leaf unit: 1
##                     x      y
## ________________________________________________
##    4            3211| 0* |11124            5
##   18   998777766665555| 0. |5667777888899 9  19
##   22            4321| 1* |011234           (6)
##   (9)       988766555| 1. |5667789          (7)
##   19      44433221000| 2* |12222333444      18
##    8         9998777| 2. |56778             7
##    1               0| 3* |00                2
## ________________________________________________
## n:                 50      50
## ________________________________________________
```

Example

Here is a real Back-to-Back Histogram Plot: Selfiecity.

Visualizing 2D Data

In case we have a 2D numerical Dataset

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n),$$

we usually do the ScaterPlot - the plot of all points (x_i, y_i), $i = 1, ...n$.

Example

Say, consider again the *cars* Dataset:

```r
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

```r
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

Example

Say, consider again the *cars* Dataset:

```
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

```
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

It has 2 Variables: *Speed* and *Distance*, and 50 Observations.

Example

Say, consider again the *cars* Dataset:

```
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

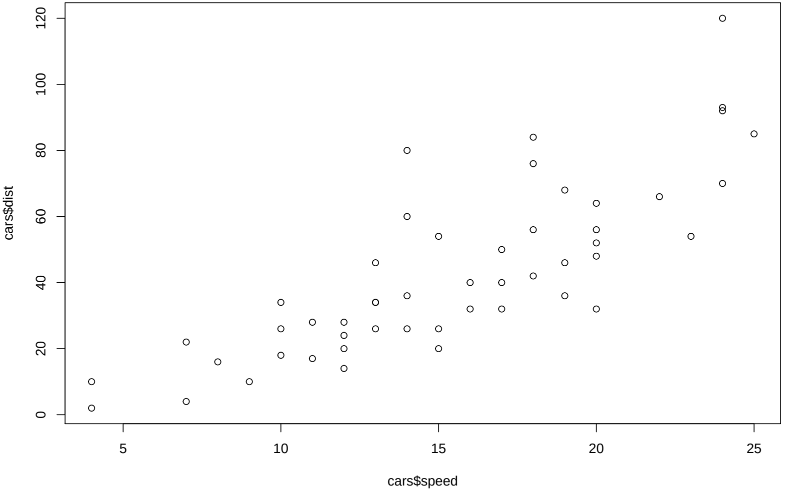
```
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

It has 2 Variables: *Speed* and *Distance*, and 50 Observations. Let us do the ScatterPlot of Observations:

ScatterPlot

```
plot(cars$speed, cars$dist)
```



Notes

- In this graph you can see that there is some relationship between the *Speed* and *Distance*, there is a *trend*: if the speed gets larger, the (stopping) distance is tending to increase.

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one.

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one.
Some ideas for multidimensional Visualizations:

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. Some ideas for multidimensional Visualizations:

- One can draw 3D in 3D $\ddot\smile$,

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. Some ideas for multidimensional Visualizations:

- One can draw 3D in 3D $\ddot\smile$, give some 3D Histograms and KDEs

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. Some ideas for multidimensional Visualizations:

- One can draw 3D in 3D $\ddot\smile$, give some 3D Histograms and KDEs
- One can draw 3D in 2D, using the 3rd variable as the Color (not in all cases, of course)

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. Some ideas for multidimensional Visualizations:

- One can draw 3D in 3D $\ddot{\smile}$, give some 3D Histograms and KDEs
- One can draw 3D in 2D, using the 3rd variable as the Color (not in all cases, of course)
- One can add the 4th Dimension by using the Size of Points

Examples

See, for example, beautiful visualizations by **Hans Rosling**.

Examples

See, for example, beautiful visualizations by **Hans Rosling**. Say, this short one: Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

Examples

See, for example, beautiful visualizations by **Hans Rosling**. Say, this short one: Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

Or, the following one: Gender Gap in Earnings per University

Additions: Multidimensional Graphs

- One can do the Pairs Plot

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the HeatMap

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the HeatMap
- One can use a Dimensionality Reduction Methods to Visualize some high dimensional Data

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the HeatMap
- One can use a Dimensionality Reduction Methods to Visualize some high dimensional Data
- etc ...

Numerical Summaries

Numerical Summaries

- Summaries (Statistics) about the Center, Mean

Numerical Summaries

- Summaries (Statistics) about the Center, Mean
- Summaries (Statistics) about the Spread

Order Statistics

First we introduce the **Order Statistics**.

Order Statistics

First we introduce the **Order Statistics**.

Assume we have a 1D Numerical Dataset $x_1, x_2, ..., x_n$.

Order Statistics

First we introduce the **Order Statistics**.

Assume we have a 1D Numerical Dataset $x_1, x_2, ..., x_n$. We sort this Dataset in the increasing order, and denote by $x_{(j)}$ the j-th element in the sorted array.

Order Statistics

First we introduce the **Order Statistics**.

Assume we have a 1D Numerical Dataset $x_1, x_2, ..., x_n$. We sort this Dataset in the increasing order, and denote by $x_{(j)}$ the j-th element in the sorted array. $x_{(j)}$ is called the j**-th Order Statistics** of our Dataset.

Order Statistics

First we introduce the **Order Statistics**.

Assume we have a 1D Numerical Dataset $x_1, x_2, ..., x_n$. We sort this Dataset in the increasing order, and denote by $x_{(j)}$ the j-th element in the sorted array. $x_{(j)}$ is called the **j-th Order Statistics** of our Dataset.

In other word, $x_{(1)}, x_{(2)}, ..., x_{(n)}$ is just a reordering of our Dataset with

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}.$$

Order Statistics

First we introduce the **Order Statistics**.

Assume we have a 1D Numerical Dataset $x_1, x_2, ..., x_n$. We sort this Dataset in the increasing order, and denote by $x_{(j)}$ the j-th element in the sorted array. $x_{(j)}$ is called the ***j*-th Order Statistics** of our Dataset.

In other word, $x_{(1)}, x_{(2)}, ..., x_{(n)}$ is just a reordering of our Dataset with

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}.$$

In particular,

$$x_{(1)} = \min\{x_1, x_2, ..., x_n\} \qquad \text{and} \qquad x_{(n)} = \max\{x_1, x_2, ..., x_n\}.$$

Example

Example: Let x be the Dataset

$$-2, 1, 3, 2, 2, 1, 1$$

Find the 4-th and 5-th Order Statistics.