AUA CS108, Statistics, Fall 2020

Lecture 07

Michael Poghosyan

9 Sep 2020

Contents

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 ScatterPlot - the plot of all points (x_i, y_i), $i = 1, ...n$.

Example

Example: Graph the ScatterPlot for the following data:

Person ID	Age	Weight
1	20	69
2	22	57
3	40	65
4	20	70

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
```

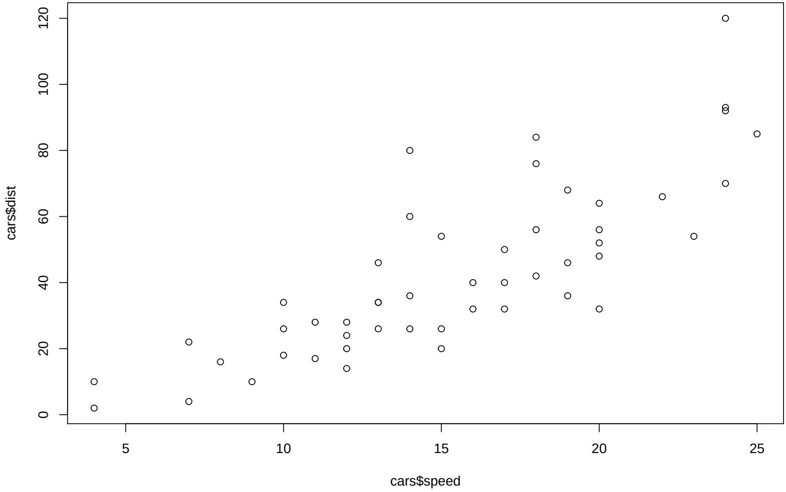
```
str(cars)
```

```
## '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. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And some ideas for multidimensional Visualizations:

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And 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. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And 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. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And 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. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And 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
- And add the 5-th one by using the Shape 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 Correlation Matrix HeatMap

Additions: Multidimensional Graphs

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

Numerical Summaries

Numerical Summaries

For 1D Datasets, we will consider the following Summaries:

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

Numerical Summaries

For 1D Datasets, we will consider the following Summaries:

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

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.

Statistical Measures for the Central Tendency/Location

Statistical Measures for the Central Tendency/Location

Here we want to answer to the questions: what are the typical values of our Dataset, where is our Data located at?

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$.

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$. We want to describe its typical value, its center.

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$. We want to describe its typical value, its center.

- **The Sample Mean:**

$$\bar{x} = mean(x) = \frac{x_1 + x_2 + ... + x_n}{n}$$

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$. We want to describe its typical value, its center.

- **The Sample Mean:**

$$\bar{x} = mean(x) = \frac{x_1 + x_2 + ... + x_n}{n}$$

Drawback: Sensitive to outliers (non-typical elements)

Sample Mean

Assume we are given a 1D numerical Dataset $x : x_1, x_2, ..., x_n$. We want to describe its typical value, its center.

- **The Sample Mean:**

$$\bar{x} = mean(x) = \frac{x_1 + x_2 + ... + x_n}{n}$$

Drawback: Sensitive to outliers (non-typical elements)

Note: Sometimes this property is a plus, not a drawback! Say, if we want to have an estimator which is sensitive to outliers.