ASDS Statistics, YSU, Fall 2020

Lecture 02

Michael Poghosyan

16 Sep 2020

Contents

Last Lecture Recap

- What is Statistics?

Last Lecture Recap

- What is Statistics?
- What is the Population?

Last Lecture Recap

- What is Statistics?
- What is the Population?
- What is a Sample?

Last Lecture Recap

- What is Statistics?
- What is the Population?
- What is a Sample?
- What is an Observation?

Last Lecture Recap

- What is Statistics?
- What is the Population?
- What is a Sample?
- What is an Observation?
- What is a Variable?

Last Lecture Recap

- What is Statistics?
- What is the Population?
- What is a Sample?
- What is an Observation?
- What is a Variable?
- What is a Parameter ?

Last Lecture Recap

- What is Statistics?
- What is the Population?
- What is a Sample?
- What is an Observation?
- What is a Variable?
- What is a Parameter ?
- What is a Statistics (again $\ddot\smile$)?

Example

Example: AMS wants to calculate the average salary for all US Mathematicians.

Example

Example: AMS wants to calculate the average salary for all US Mathematicians.

Can you describe

- the **Population**

Example

Example: AMS wants to calculate the average salary for all US Mathematicians.

Can you describe

- the **Population**
- a **Sample**

Example

Example: AMS wants to calculate the average salary for all US Mathematicians.

Can you describe

- the **Population**
- a **Sample**
- the **Parameter**

Example

Example: AMS wants to calculate the average salary for all US Mathematicians.

Can you describe

- the **Population**
- a **Sample**
- the **Parameter**
- the **Statistics**

Example

Example: AMS wants to calculate the average salary for all US Mathematicians.

Can you describe

- the **Population**
- a **Sample**
- the **Parameter**
- the **Statistics**
- an **Observation** ?

Collecting the Data

If you want to get some trustworthy information, make reliable generalizations and good predictions from your Data, your Data need to be a **good** one.

Collecting the Data

If you want to get some trustworthy information, make reliable generalizations and good predictions from your Data, your Data need to be a **good** one.

First, for doing Statistics, Statisticians are modeling the process of Data Collection, they are *Designing the Experiment and the Sampling Methodology*.

Collecting the Data

If you want to get some trustworthy information, make reliable generalizations and good predictions from your Data, your Data need to be a **good** one.

First, for doing Statistics, Statisticians are modeling the process of Data Collection, they are *Designing the Experiment and the Sampling Methodology*. Correct design is very important for doing a correct analysis.

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies $\smile$).

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies $\ddot\smile$). Well, we cannot ask *every* person in Armenia.

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies $\ddot\smile$). Well, we cannot ask *every* person in Armenia. Instead, on one Friday, from 9AM till 6PM, we stand in front of the entrance of the "Marshal Baghramyan" metro station and ask every person we meet about his/her English knowledge.

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies $\ddot\smile$). Well, we cannot ask *every* person in Armenia. Instead, on one Friday, from 9AM till 6PM, we stand in front of the entrance of the "Marshal Baghramyan" metro station and ask every person we meet about his/her English knowledge.

Is this a good choice of a Sample?

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies $\ddot\smile$). Well, we cannot ask *every* person in Armenia. Instead, on one Friday, from 9AM till 6PM, we stand in front of the entrance of the "Marshal Baghramyan" metro station and ask every person we meet about his/her English knowledge.

Is this a good choice of a Sample? What is wrong here?

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies ☺). Well, we cannot ask *every* person in Armenia. Instead, on one Friday, from 9AM till 6PM, we stand in front of the entrance of the "Marshal Baghramyan" metro station and ask every person we meet about his/her English knowledge.

Is this a good choice of a Sample? What is wrong here?

Example: There are different (real) examples from older days of wrong conclusions made by using exit polls about the (presidential) elections in USA.

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies $\ddot\smile$). Well, we cannot ask *every* person in Armenia. Instead, on one Friday, from 9AM till 6PM, we stand in front of the entrance of the "Marshal Baghramyan" metro station and ask every person we meet about his/her English knowledge.

Is this a good choice of a Sample? What is wrong here?

Example: There are different (real) examples from older days of wrong conclusions made by using exit polls about the (presidential) elections in USA. Say, one of the very respective newspapers made an exit poll by randomly calling by a phone its subscribers and asking about their choice.

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies $\ddot\smile$). Well, we cannot ask *every* person in Armenia. Instead, on one Friday, from 9AM till 6PM, we stand in front of the entrance of the "Marshal Baghramyan" metro station and ask every person we meet about his/her English knowledge.

Is this a good choice of a Sample? What is wrong here?

Example: There are different (real) examples from older days of wrong conclusions made by using exit polls about the (presidential) elections in USA. Say, one of the very respective newspapers made an exit poll by randomly calling by a phone its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite.

Examples: Biased Sampling

Example: Assume we want to get information about the ratio of English-speaking persons in Armenia (who can speak, of course, not babies $\ddot\smile$). Well, we cannot ask *every* person in Armenia. Instead, on one Friday, from 9AM till 6PM, we stand in front of the entrance of the "Marshal Baghramyan" metro station and ask every person we meet about his/her English knowledge.

Is this a good choice of a Sample? What is wrong here?

Example: There are different (real) examples from older days of wrong conclusions made by using exit polls about the (presidential) elections in USA. Say, one of the very respective newspapers made an exit poll by randomly calling by a phone its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite. Guess why?

Examples: Biased Sampling

Example: Assume the ad says: *91% of customers choose our shampoo "Voskemazik"*.

Examples: Biased Sampling

Example: Assume the ad says: *91% of customers choose our shampoo "Voskemazik"*.

Can this be true?

Examples: Biased Sampling

Example: Assume the ad says: *91% of customers choose our shampoo "Voskemazik"*.

Can this be true? Can this be true but give wrong information?

Examples: Biased Sampling

Example: Assume the ad says: *91% of customers choose our shampoo "Voskemazik"*.

Can this be true? Can this be true but give wrong information?

Example: Recall the Experiment to calculate the proportion of female students in YSU.

Examples: Biased Sampling

Example: Assume the ad says: *91% of customers choose our shampoo "Voskemazik"*.

Can this be true? Can this be true but give wrong information?

Example: Recall the Experiment to calculate the proportion of female students in YSU.

Is that Sample representative?

Examples: Biased Sampling

Example: Assume the ad says: *91% of customers choose our shampoo "Voskemazik"*.

Can this be true? Can this be true but give wrong information?

Example: Recall the Experiment to calculate the proportion of female students in YSU.

Is that Sample representative? Why?

Random Sampling

The moral is that for correct Statistical Analysis, one needs to have a Correct Data.

Random Sampling

The moral is that for correct Statistical Analysis, one needs to have a Correct Data. If your Data is Biased, you will never get correct information.

Random Sampling

The moral is that for correct Statistical Analysis, one needs to have a Correct Data. If your Data is Biased, you will never get correct information.

(Un)fortunately, we will not go into the details of the Experimental and Sampling Design. From this point on we will assume that we have a **Representative Sample**, obtained through a Simple Random Sampling:

Random Sampling

The moral is that for correct Statistical Analysis, one needs to have a Correct Data. If your Data is Biased, you will never get correct information.

(Un)fortunately, we will not go into the details of the Experimental and Sampling Design. From this point on we will assume that we have a **Representative Sample**, obtained through a Simple Random Sampling: Say, we want to have a Sample of size (number of elements) k.

Random Sampling

The moral is that for correct Statistical Analysis, one needs to have a Correct Data. If your Data is Biased, you will never get correct information.

(Un)fortunately, we will not go into the details of the Experimental and Sampling Design. From this point on we will assume that we have a **Representative Sample**, obtained through a Simple Random Sampling: Say, we want to have a Sample of size (number of elements) k.

Definition: We say that our Sample is *Representative* (obtained by a Simple Random Sampling), if it is obtained in the process where all Samples of size k have the same probability of being chosen.

Example

Example: Assume we have 10 male and 20 female students in our class, and we want to choose a sample of size 6. Here are some possibilities:

Example

Example: Assume we have 10 male and 20 female students in our class, and we want to choose a sample of size 6. Here are some possibilities:

- Choose at random 2 male students and 4 female students;

Example

Example: Assume we have 10 male and 20 female students in our class, and we want to choose a sample of size 6. Here are some possibilities:

- Choose at random 2 male students and 4 female students;
- Choose at random 3 male and 3 female students;

Example

Example: Assume we have 10 male and 20 female students in our class, and we want to choose a sample of size 6. Here are some possibilities:

- Choose at random 2 male students and 4 female students;
- Choose at random 3 male and 3 female students;
- Choose at random 6 names from the list of all 30 students

Example

Example: Assume we have 10 male and 20 female students in our class, and we want to choose a sample of size 6. Here are some possibilities:

- Choose at random 2 male students and 4 female students;
- Choose at random 3 male and 3 female students;
- Choose at random 6 names from the list of all 30 students

Which one gives a Representative (Simple Random) Sample?

Few Sampling Methods

Simple Random Sampling is not always easy to perform, so people are using different simpler Sampling Strategies (although they are not always giving exactly Representative Samples):

Few Sampling Methods

- *Systematic (Interval) Sampling*, we fit the population into a list, enumerate it; choose n; choose at random a starting element from the first n elements in the list; and from that element on, every n-th member of the population is selected;

Few Sampling Methods

- *Systematic (Interval) Sampling*, we fit the population into a list, enumerate it; choose n; choose at random a starting element from the first n elements in the list; and from that element on, every n-th member of the population is selected;
- *Stratified Sampling*, where the total population is divided into subgroups (strata), that share similar characteristics, and then random sample (of corresponding size) is chosen from each strata.

Few Sampling Methods

- *Systematic (Interval) Sampling*, we fit the population into a list, enumerate it; choose n; choose at random a starting element from the first n elements in the list; and from that element on, every n-th member of the population is selected;
- *Stratified Sampling*, where the total population is divided into subgroups (strata), that share similar characteristics, and then random sample (of corresponding size) is chosen from each strata.
- *Cluster Sampling*, where the total population is divided into subgroups (clusters), then some clusters are randomly chosen. Then we include all elements of chosen clusters into our Sample.

Classification of Data wrt its Dimension

Data can be

- **Univariate** (1D) - here the observations are on a single Variable
- **Bivariate** (2D) - here the observations are on two Variables
- **Multivariate** (n-D, $n \geq 2$) - when the observations are on more than a one Variable (usually, more than two)

Classification of Variables wrt its Type

We need this classification/differentiation, because each type requires different Statistical approaches for an analysis.

Classification of Variables wrt its Type

We need this classification/differentiation, because each type requires different Statistical approaches for an analysis.

In Statistics, we deal with 2 types of Variables:

Classification of Variables wrt its Type

We need this classification/differentiation, because each type requires different Statistical approaches for an analysis.

In Statistics, we deal with 2 types of Variables:

- **Qualitative** or **Categorical** Variable: the value is a category, non-numerical

Classification of Variables wrt its Type

We need this classification/differentiation, because each type requires different Statistical approaches for an analysis.

In Statistics, we deal with 2 types of Variables:

- **Qualitative** or **Categorical** Variable: the value is a category, non-numerical

Examples: *Gender* is a Categorical Variable, taking values *Male, Female*;

Classification of Variables wrt its Type

We need this classification/differentiation, because each type requires different Statistical approaches for an analysis.

In Statistics, we deal with 2 types of Variables:

- **Qualitative** or **Categorical** Variable: the value is a category, non-numerical

Examples: *Gender* is a Categorical Variable, taking values *Male, Female*; Or *Color* and *Model* (of a car) are again Categorical

Classification of Variables wrt its Type

- **Quantitative** or **Numerical** Variable: the value is a number obtained from counting, measuring something etc.

Classification of Variables wrt its Type

- **Quantitative** or **Numerical** Variable: the value is a number obtained from counting, measuring something etc. In this case we differentiate between
  - **Discrete:** the range is finite or countably infinite

Classification of Variables wrt its Type

- **Quantitative** or **Numerical** Variable: the value is a number obtained from counting, measuring something etc. In this case we differentiate between
  - **Discrete:** the range is finite or countably infinite
  - **Continuous:** the range is some interval

Classification of Variables wrt its Type

- **Quantitative** or **Numerical** Variable: the value is a number obtained from counting, measuring something etc. In this case we differentiate between
  - **Discrete:** the range is finite or countably infinite
  - **Continuous:** the range is some interval

Examples:

- *No. of Children*, *No. of Customers* ,... are Discrete

Classification of Variables wrt its Type

- **Quantitative** or **Numerical** Variable: the value is a number obtained from counting, measuring something etc. In this case we differentiate between
  - **Discrete:** the range is finite or countably infinite
  - **Continuous:** the range is some interval

Examples:

- *No. of Children*, *No. of Customers* ,... are Discrete
- *Height*, *Weight*, *Age*, ... are Continuous

Remark

Remark: Of course, we can enumerate Categorical Data, say, instead of *Male, Female* we can just use 0 and 1. It seems that we have already a Numerical, Quantitative (Discrete) Data. But there is a difference:

Remark

Remark: Of course, we can enumerate Categorical Data, say, instead of *Male, Female* we can just use 0 and 1. It seems that we have already a Numerical, Quantitative (Discrete) Data. But there is a difference:

Let me give by an example: when talking about the number of children in the family, we can have the following data: 0, 2, 1, 2, 4, 6, and we can calculate, say, the average number of children in families, here 2.5.

Remark

Remark: Of course, we can enumerate Categorical Data, say, instead of *Male, Female* we can just use 0 and 1. It seems that we have already a Numerical, Quantitative (Discrete) Data. But there is a difference:

Let me give by an example: when talking about the number of children in the family, we can have the following data: 0, 2, 1, 2, 4, 6, and we can calculate, say, the average number of children in families, here 2.5.

But even if we are enumerating the Gender or the Color, the average Gender or the average Color is not meaningful, we cannot deal with the assigned numbers as above!

Remark

Remark: Categorical variables also can be divided into two types: Nominal and Ordinal.

Remark

Remark: Categorical variables also can be divided into two types: Nominal and Ordinal.

- **Nominal** - This is when we assign categories to observations, and categories do not have a logical order.

Remark

Remark: Categorical variables also can be divided into two types: Nominal and Ordinal.

- **Nominal** - This is when we assign categories to observations, and categories do not have a logical order.

Say, the Variable *Gender* can take values “male”, “female”; or the Variables *Marital Status*; *blood group*; *Mobile Phone Manufacturer*; . . .

Remark

Remark: Categorical variables also can be divided into two types: Nominal and Ordinal.

- **Nominal** - This is when we assign categories to observations, and categories do not have a logical order.

Say, the Variable *Gender* can take values "male", "female"; or the Variables *Marital Status*; *blood group*; *Mobile Phone Manufacturer*; ...

- **Ordinal** - This is when we assign categories to observations, but this time we have an intrinsic order.

Remark

Remark: Categorical variables also can be divided into two types: Nominal and Ordinal.

- **Nominal** - This is when we assign categories to observations, and categories do not have a logical order.

Say, the Variable *Gender* can take values “male”, “female”; or the Variables *Marital Status*; *blood group*; *Mobile Phone Manufacturer*; ...

- **Ordinal** - This is when we assign categories to observations, but this time we have an intrinsic order. Say, *Year of study* (freshman, sopho,...), *Letter Grade* or *Education* (HS, BS, MS, PhD) are on the Ordinal Scale

Remark

Remark: Categorical variables also can be divided into two types: Nominal and Ordinal.

- **Nominal** - This is when we assign categories to observations, and categories do not have a logical order.

Say, the Variable *Gender* can take values “male”, “female”; or the Variables *Marital Status*; *blood group*; *Mobile Phone Manufacturer*; ...

- **Ordinal** - This is when we assign categories to observations, but this time we have an intrinsic order. Say, *Year of study* (freshman, sopho,...), *Letter Grade* or *Education* (HS, BS, MS, PhD) are on the Ordinal Scale

Maybe one of the well-known Ordinal Scale Measurements is the **Likert Scale**:

Remark

Remark: Categorical variables also can be divided into two types: Nominal and Ordinal.

- **Nominal** - This is when we assign categories to observations, and categories do not have a logical order.

Say, the Variable *Gender* can take values "male", "female"; or the Variables *Marital Status*; *blood group*; *Mobile Phone Manufacturer*; ...

- **Ordinal** - This is when we assign categories to observations, but this time we have an intrinsic order. Say, *Year of study* (freshman, sopho,...), *Letter Grade* or *Education* (HS, BS, MS, PhD) are on the Ordinal Scale

Maybe one of the well-known Ordinal Scale Measurements is the **Likert Scale**: This is our famous

Strongly Disagree | Disagree | Neither | Agree | Strongly Agree

Descriptive Statistics

Why we need it?

Descriptive Statistics

Descriptive Statistics is to get the first, basic information about the Data, either in the Visual or Numerical form.

Descriptive Statistics

Descriptive Statistics is to get the first, basic information about the Data, either in the Visual or Numerical form.

Consider, for example, the dataset `mpg` from the `ggplot2` package.

Descriptive Statistics

Descriptive Statistics is to get the first, basic information about the Data, either in the Visual or Numerical form.

Consider, for example, the dataset `mpg` from the `ggplot2` package. From the official description of the data,

> *This dataset contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008.*

Descriptive Statistics

Descriptive Statistics is to get the first, basic information about the Data, either in the Visual or Numerical form.

Consider, for example, the dataset `mpg` from the `ggplot2` package. From the official description of the data,

> *This dataset contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008.*

Lets look at the first 3 rows of our dataset:

```r
head(ggplot2::mpg, 3)
```

```
## # A tibble: 3 x 11
##   manufacturer model displ  year   cyl trans      drv     cty
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18
## 2 audi         a4      1.8  1999     4 manual(m5) f        21
## 3 audi         a4      2    2008     4 manual(m6) f        20
```

Descriptive Statistics

Descriptive Statistics

The variable `cty` is the *city miles per gallon*, and the variable `cyl` is the *number of cylinders*. Let's separate that Variables:

```
cty <- ggplot2::mpg$cty
cyl <- ggplot2::mpg$cyl
```

Descriptive Statistics

Let's see the results:

```
cyl
```

```
##   [1] 4 4 4 4 6 6 6 4 4 4 4 6 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8
##  [38] 4 6 6 6 6 6 6 6 6 6 6 6 6 6 6 8 8 8 8 8 6 8 8 8 8 8 8 8
##  [75] 8 8 8 6 6 6 6 8 8 6 6 8 8 8 8 8 6 6 6 6 8 8 8 8 8 4 4 4
## [112] 4 6 6 6 4 4 4 4 6 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8 8 8 6 6
## [149] 6 6 6 6 6 8 6 6 6 6 8 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6
## [186] 6 4 4 4 4 6 6 6 4 4 4 4 4 8 8 4 4 4 6 6 6 6 4 4 4 4 6 4
## [223] 4 4 4 5 5 4 4 4 4 6 6 6
```

Descriptive Statistics

Let's see the results:

```r
cyl
```

```
##   [1] 4 4 4 4 6 6 6 4 4 4 4 6 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8 8
##  [38] 4 6 6 6 6 6 6 6 6 6 6 6 6 6 6 8 8 8 8 8 6 8 8 8 8 8 8 8
##  [75] 8 8 8 6 6 6 6 8 8 6 6 8 8 8 8 8 6 6 6 6 8 8 8 8 8 4 4 4
## [112] 4 6 6 6 4 4 4 4 6 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8 8 8 6 6
## [149] 6 6 6 6 6 8 6 6 6 6 8 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6
## [186] 6 4 4 4 4 6 6 6 4 4 4 4 4 8 8 4 4 4 6 6 6 6 4 4 4 4 6 4
## [223] 4 4 4 5 5 4 4 4 4 6 6 6
```

Can you describe this data? What can be said about the No. of Cylinders of these cars?

Descriptive Statistics

Let's see the results for `cty`:

```
cty
```

```
##   [1] 18 21 20 21 16 18 18 18 16 20 19 15 17 17 15 15 17 16 1
##  [26] 16 15 15 14 11 11 14 19 22 18 18 17 18 17 16 16 17 17 1
##  [51] 13 14 14 14  9 11 11 13 13  9 13 11 13 11 12  9 13 13 1
##  [76] 11 12 14 15 14 13 13 13 14 14 13 13 13 11 13 18 18 17 1
## [101] 24 25 23 24 26 25 24 21 18 18 21 21 18 18 19 19 19 20 2
## [126] 14  9 14 13 11 11 12 12 11 11 11 12 14 13 13 13 21 19 2
## [151] 14 15 14 12 18 16 17 18 16 18 18 20 19 20 18 21 19 19 1
## [176] 15 15 16 14 21 21 21 21 18 18 19 21 21 21 22 18 18 18 2
## [201] 15 16 17 15 15 15 16 21 19 21 22 17 33 21 19 22 21 21 2
## [226] 20 20 21 18 19 21 16 18 17
```

Descriptive Statistics

Let's see the results for `cty`:

```
cty
```

```
##   [1] 18 21 20 21 16 18 18 18 16 20 19 15 17 17 15 15 17 16 1
##  [26] 16 15 15 14 11 11 14 19 22 18 18 17 18 17 16 16 17 17 1
##  [51] 13 14 14 14  9 11 11 13 13  9 13 11 13 11 12  9 13 13 1
##  [76] 11 12 14 15 14 13 13 13 14 14 13 13 13 11 13 18 18 17 1
## [101] 24 25 23 24 26 25 24 21 18 18 21 21 18 18 19 19 19 20 2
## [126] 14  9 14 13 11 11 12 12 11 11 11 12 14 13 13 13 21 19 2
## [151] 14 15 14 12 18 16 17 18 16 18 18 20 19 20 18 21 19 19 1
## [176] 15 15 16 14 21 21 21 21 18 18 19 21 21 21 22 18 18 18 2
## [201] 15 16 17 15 15 15 16 21 19 21 22 17 33 21 19 22 21 21 2
## [226] 20 20 21 18 19 21 16 18 17
```

Again, can you describe this data? What can be said about the City Miles per Gallon values of these cars?

Descriptive Statistics

Descriptive Statistics gives us tools to Describe the Data, get some basic, general information about the dataset.

Descriptive Statistics

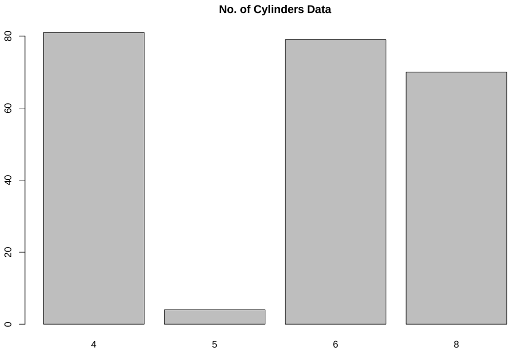
Descriptive Statistics gives us tools to Describe the Data, get some basic, general information about the dataset. We will describe data Graphically and/or by giving some numerical summaries.

Descriptive Statistics

Descriptive Statistics gives us tools to Describe the Data, get some basic, general information about the dataset. We will describe data Graphically and/or by giving some numerical summaries.

For example, let us draw the BarPlot for the frequencies of the `cyl` variable:

```
barplot(table(cyl), main = "No. of Cylinders Data")
```



Descriptive Statistics

Now, let us give some numerical summaries for `cty`: calculate the average Miles per Gallon for a City, and its max and min.

```r
cat("mean = ", mean(cty))
```

```
## mean =  16.85897
```

```r
cat("Max = ", max(cty))
```

```
## Max =  35
```

```r
cat("Min = ", min(cty))
```

```
## Min =  9
```

Descriptive Statistics

Now, let us give some numerical summaries for `cty`: calculate the average Miles per Gallon for a City, and its max and min.

```
cat("mean = ", mean(cty))
```

```
## mean =  16.85897
```

```
cat("Max = ", max(cty))
```

```
## Max =  35
```

```
cat("Min = ", min(cty))
```

```
## Min =  9
```

And we can use the `summary` command to get some numerical info:

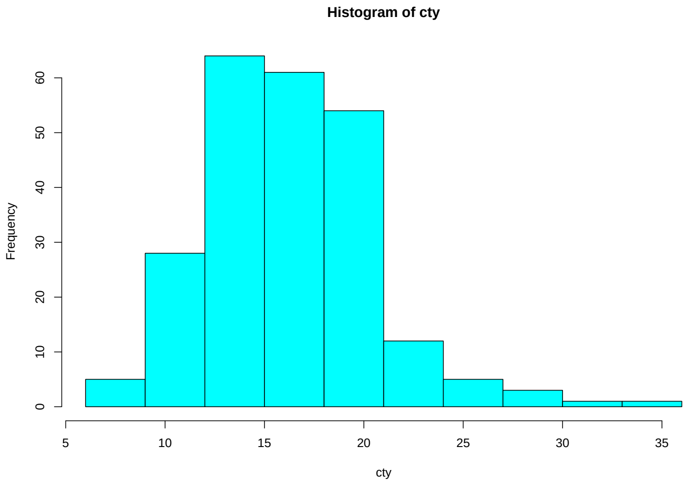
```
summary(cty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   14.00   17.00   16.86   19.00   35.00
```

Descriptive Statistics

To get some visual information about the Variable `cty`, its distribution, we can draw the Histogram:

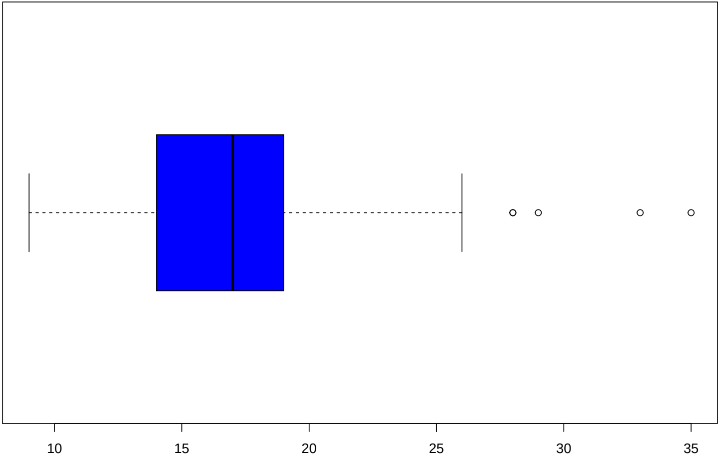
```
hist(cty, breaks=seq(6,36, 3),col="cyan")
```



Descriptive Statistics

Now, we can draw the BoxPlot of the `cty` data:

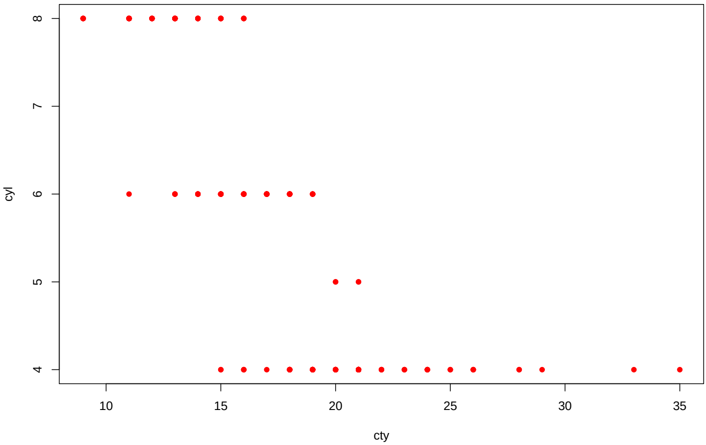
```
boxplot(cty, horizontal = T, col = "blue")
```



Descriptive Statistics

Now, instead of just getting information about `cyl` and `cty` separately, let us give visually the relationship between them:

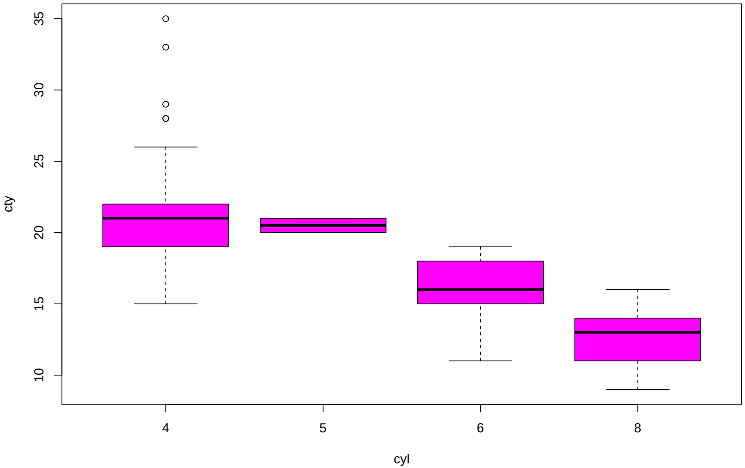
```
plot(cty, cyl, pch=16, col = "red")
```



Descriptive Statistics

... or draw a BoxPlot of `cty` for each type of the cylinder:

```
boxplot(cty~cyl, col="magenta")
```



Descriptive Statistics

Moral: our brain cannot get an insight from the list of numbers, but Descriptive Statistics can help $\ddot\smile$

Descriptive Statistics

How to do it?

Descriptive Statistics for a Univariate Data

So we use the Descriptive Statistics at the beginning of our Statistical Study to examine, explore the dataset.

Descriptive Statistics for a Univariate Data

So we use the Descriptive Statistics at the beginning of our Statistical Study to examine, explore the dataset.

This part of Statistics is sometimes called **Exploratory Data Analysis**, EDA.

Descriptive Statistics for a Univariate Data

So we use the Descriptive Statistics at the beginning of our Statistical Study to examine, explore the dataset.

This part of Statistics is sometimes called **Exploratory Data Analysis**, EDA.

And we start by describing some of the *Graphical Summaries*.

Descriptive Statistics for a Univariate Data

So we use the Descriptive Statistics at the beginning of our Statistical Study to examine, explore the dataset.

This part of Statistics is sometimes called **Exploratory Data Analysis**, EDA.

And we start by describing some of the *Graphical Summaries*.

Here, for the beginning, we will assume that we have a univariate (mostly numerical) data (dataset), $x_1, x_2, ..., x_n$. In this case we will say that we are given a (univariate, 1D) dataset x.

Frequency Tables

Here we assume that we have observations from a 1D numerical or categorical variable, i.e., we have a univariate *discrete* numerical or categorical data $x_1, x_2, ..., x_n$.

Frequency Tables

Here we assume that we have observations from a 1D numerical or categorical variable, i.e., we have a univariate *discrete* numerical or categorical data $x_1, x_2, ..., x_n$.

Definition: The **frequency** of a value t in observations $x_1, x_2, ..., x_n$ is the number of times t occurs in observations:

$$\text{Frequency of } t = \text{number of occurrences of } t \text{ in data.}$$

Frequency Tables

Here we assume that we have observations from a 1D numerical or categorical variable, i.e., we have a univariate *discrete* numerical or categorical data $x_1, x_2, ..., x_n$.

Definition: The **frequency** of a value t in observations $x_1, x_2, ..., x_n$ is the number of times t occurs in observations:

$$\text{Frequency of } t = \text{number of occurrences of } t \text{ in data.}$$

Definition: The **relative frequency** (or percentage) of a value t in observations $x_1, x_2, ..., x_n$ is the ratio of frequency of t divided by the total number of observations, n:

$$\text{Relative Frequency of } t = \frac{\text{Frequency of } t}{\text{Total Number of Observations}} =$$

$$= \frac{\text{Frequency of } t}{n}.$$

Frequency Tables, Example

Example: Given the following Dataset:

$$1, 2, 4, 7, 2, 3, 2, 1, 2, 1, 4, 1, -1$$

obtain the Frequency and Relative Frequency Tables.

Frequency Tables, Example

Example: Given the following Dataset:

$$1, 2, 4, 7, 2, 3, 2, 1, 2, 1, 4, 1, -1$$

obtain the Frequency and Relative Frequency Tables.

Example: Let's construct the Frequency Table of the above Dataset using **R**:

```r
x <- c(1, 2, 4, 7, 2, 3, 2, 1, 2, 1, 4, 1, -1)
table(x)
```

```
## x
## -1  1  2  3  4  7
##  1  4  4  1  2  1
```

Cumulative Frequency and Relative Frequency Tables

For tabular representation of Discrete Numerical Data, people are sometimes using Cumulative Frequency and Cumulative Relative Frequency Tables:

Cumulative Frequency and Relative Frequency Tables

For tabular representation of Discrete Numerical Data, people are sometimes using Cumulative Frequency and Cumulative Relative Frequency Tables:

Example: Form the Cumulative Frequency and Relative Frequency tables for the following data:

$$1, 2, 4, 1, 1, 2.$$

Cumulative Frequency and Relative Frequency Tables

For tabular representation of Discrete Numerical Data, people are sometimes using Cumulative Frequency and Cumulative Relative Frequency Tables:

Example: Form the Cumulative Frequency and Relative Frequency tables for the following data:

$$1, 2, 4, 1, 1, 2.$$

We will meet, in fact, the Cumulative Relative Frequency Table soon, under the name Empirical CDF, ECDF.