YSU ASDS, Statistics, Fall 2019

Lectures 02 - 03

Michael Poghosyan

04 Sep 2019

Descriptive Statistics

Contents

Last Lecture ReCap

- What is a **Population**?

Last Lecture ReCap

- What is a **Population**?
- What is a **Sample**?

Last Lecture ReCap

- What is a **Population**?
- What is a **Sample**?
- Can you give an example of a Statistical Question/Problem, describe the Population and a Sample?

Last Lecture ReCap

- What is a **Population**?
- What is a **Sample**?
- Can you give an example of a Statistical Question/Problem, describe the Population and a Sample?
- What is an **Observation**?

Last Lecture ReCap

- What is a **Population**?
- What is a **Sample**?
- Can you give an example of a Statistical Question/Problem, describe the Population and a Sample?
- What is an **Observation**?
- What is a **Variable**?

Last Lecture ReCap

- What is a **Population**?
- What is a **Sample**?
- Can you give an example of a Statistical Question/Problem, describe the Population and a Sample?
- What is an **Observation**?
- What is a **Variable**?
- What is a **Parameter**?

Last Lecture ReCap

- What is a **Population**?
- What is a **Sample**?
- Can you give an example of a Statistical Question/Problem, describe the Population and a Sample?
- What is an **Observation**?
- What is a **Variable**?
- What is a **Parameter**?
- What is a **Statistics** (not the Subject!)?

Last Lecture ReCap

- What is a **Population**?
- What is a **Sample**?
- Can you give an example of a Statistical Question/Problem, describe the Population and a Sample?
- What is an **Observation**?
- What is a **Variable**?
- What is a **Parameter**?
- What is a **Statistics** (not the Subject!)?
- What is a **Representative Sample**?

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

- Collecting the Data

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

- Collecting the Data
  - Processing Data: Organizing, Cleaning, Curating, ...

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

- Collecting the Data
  - Processing Data: Organizing, Cleaning, Curating, ...
- Visualizing/Describing of the Data

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

- Collecting the Data
  - Processing Data: Organizing, Cleaning, Curating, ...
- Visualizing/Describing of the Data
- Doing a Statistical Analysis and Inference

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

- Collecting the Data
  - Processing Data: Organizing, Cleaning, Curating, ...
- Visualizing/Describing of the Data
- Doing a Statistical Analysis and Inference
- Drawing Conclusions, Making Predictions

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 modelling 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 modelling 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

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

Examples

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

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

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

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

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: Assume we want to study which kind of music is affecting much in faster learning to calculate Integrals $\ddot\smile$

Examples

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: Assume we want to study which kind of music is affecting much in faster learning to calculate Integrals $\ddot\smile$ To this end, we choose some group of freshmen, give them integrals, and ask them to solve the first one under Classical Rock, the second one - under Heavy Metal, the third one - under Classical Music, the next one - under Jazz, Blues, Funk, Popsa, Rabiz ect.

Examples

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: Assume we want to study which kind of music is affecting much in faster learning to calculate Integrals ⌣̈ To this end, we choose some group of freshmen, give them integrals, and ask them to solve the first one under Classical Rock, the second one - under Heavy Metal, the third one - under Classical Music, the next one - under Jazz, Blues, Funk, Popsa, Rabiz ect. And we ask them to fix the time of calculation.

Examples

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: Assume we want to study which kind of music is affecting much in faster learning to calculate Integrals ⌣̈ To this end, we choose some group of freshmen, give them integrals, and ask them to solve the first one under Classical Rock, the second one - under Heavy Metal, the third one - under Classical Music, the next one - under Jazz, Blues, Funk, Popsa, Rabiz ect. And we ask them to fix the time of calculation. Then we analyze the data, and make decisions.

Is this a good design of the Experiment?

Examples

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: Assume we want to study which kind of music is affecting much in faster learning to calculate Integrals $\ddot\smile$ To this end, we choose some group of freshmen, give them integrals, and ask them to solve the first one under Classical Rock, the second one - under Heavy Metal, the third one - under Classical Music, the next one - under Jazz, Blues, Funk, Popsa, Rabiz ect. And we ask them to fix the time of calculation. Then we analyze the data, and make decisions.

Is this a good design of the Experiment? What is wrong here?

Examples: Biased Sampling

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: 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 its subscribers and asking about their choice.

Examples: Biased Sampling

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 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: 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 its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite. Why?

Examples: Biased Sampling

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 its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite. Why?

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

Examples: Biased Sampling

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 its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite. Why?

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

Can this be true?

Examples: Biased Sampling

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 its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite. Why?

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: 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 its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite. Why?

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 ratio of female students in YSU.

Examples: Biased Sampling

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 its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite. Why?

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 ratio of female students in YSU.

Is that Sample representative?

Examples: Biased Sampling

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 its subscribers and asking about their choice. Newspaper made a conclusion from the data, but the actual result was exactly the opposite. Why?

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 ratio of female students in YSU.

Is that Sample representative? Why?

Random Sampling

The moral of the above examples is that for correct Statistical Analysis, one needs to design the Experiment wisely.

Random Sampling

The moral of the above examples is that for correct Statistical Analysis, one needs to design the Experiment wisely. (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 of the above examples is that for correct Statistical Analysis, one needs to design the Experiment wisely. (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 of the above examples is that for correct Statistical Analysis, one needs to design the Experiment wisely. (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 studens;

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 studens;
- 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 studens;
- 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 studens;
- 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 Simple Random Sample?

Few Sampling Methods

Let us recall the Definition of the **Representative Sample** (of size 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.

Few Sampling Methods

Let us recall the Definition of the **Representative Sample** (of size 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.

Simple Random Sampling is not so 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 choosen 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 choosen 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

In Statistics, we deal with 2 types of Variables:

Classification of Variables wrt its Type

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

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

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

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

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

Classification of Variables wrt its Type

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

- **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

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

- **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

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

- **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

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

- **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 Sex or the Color, the average Sex or the average Color is not meaningful, we cannot deal with the assigned numbers as above!

Classification of Data wrt Measurement Levels

Statistician differentiate between the Measurement Levels, make difference between different types of, say, Categorical Data.

Classification of Data wrt Measurement Levels

Statistician differentiate between the Measurement Levels, make difference between different types of, say, Categorical Data. This is because there are different Statistical approaches for each level.

Classification of Data wrt Measurement Levels

Statistician differentiate between the Measurement Levels, make difference between different types of, say, Categorical Data. This is because there are different Statistical approaches for each level.

Example: *Color* and *Stat Final Letter Grade* are Categorical Variables.

Classification of Data wrt Measurement Levels

Statistician differentiate between the Measurement Levels, make difference between different types of, say, Categorical Data. This is because there are different Statistical approaches for each level.

Example: *Color* and *Stat Final Letter Grade* are Categorical Variables. Or, *Sex* and *Year of University Study* (freshman, sophomore, junior, senior) are again Categorical.

Classification of Data wrt Measurement Levels

Statistician differentiate between the Measurement Levels, make difference between different types of, say, Categorical Data. This is because there are different Statistical approaches for each level.

Example: *Color* and *Stat Final Letter Grade* are Categorical Variables. Or, *Sex* and *Year of University Study* (freshman, sophomore, junior, senior) are again Categorical. What do you think, is there an essential difference between these Variables?

Classification of Data wrt Measurement Levels

Statistician differentiate between the Measurement Levels, make difference between different types of, say, Categorical Data. This is because there are different Statistical approaches for each level.

Example: *Color* and *Stat Final Letter Grade* are Categorical Variables. Or, *Sex* and *Year of University Study* (freshman, sophomore, junior, senior) are again Categorical. What do you think, is there an essential difference between these Variables?

Yeah, there is an **order** in the second Variables, *Stat Final Letter Grade* and *Year of University Study*.

Measurement Levels

The (commonly agreed) Measurement Levels are:

Measurement Levels

The (commonly agreed) Measurement Levels are:

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

Measurement Levels

The (commonly agreed) Measurement Levels are:

- **Nominal Measurements** - 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*; . . .

Measurement Levels

The (commonly agreed) Measurement Levels are:

- **Nominal Measurements** - 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 Measurements** - This is when we assign categories to observations, but this time we have an intristic order.

Measurement Levels

The (commonly agreed) Measurement Levels are:

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

Measurement Levels

The (commonly agreed) Measurement Levels are:

- **Nominal Measurements** - 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 Measurements** - This is when we assign categories to observations, but this time we have an intristic 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 Measurments is the **Likert Scale**:

Measurement Levels

The (commonly agreed) Measurement Levels are:

- **Nominal Measurements** - 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 Measurements** - This is when we assign categories to observations, but this time we have an intristic 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 Measurments is the **Likert Scale**: This is our famous

Strongly Disagree | Disagree | Neither | Agree | Strongly Agree

Measurement Levels

- **Interval Measurements** - Are like Ordinal Measurements, but here the differences are meaningful, equal differences between the category levels mean equal differences between the characteristics measured. Also here zero is arbitrary (does not mean the absence of the characteristic), and ratios do not make sense.

Measurement Levels

- **Interval Measurements** - Are like Ordinal Measurements, but here the differences are meaningful, equal differences between the category levels mean equal differences between the characteristics measured. Also here zero is arbitrary (does not mean the absence of the characteristic), and ratios do not make sense. For example, in the Likert Scale, it is not clear if the difference between *Disagree* and *Strongly Disagree* is the same as the difference between *Strongly Agree* and *Agree*.

Measurement Levels

- **Interval Measurements** - Are like Ordinal Measurements, but here the differences are meaningful, equal differences between the category levels mean equal differences between the characteristics measured. Also here zero is arbitrary (does not mean the absence of the characteristic), and ratios do not make sense. For example, in the Likert Scale, it is not clear if the difference between *Disagree* and *Strongly Disagree* is the same as the difference between *Strongly Agree* and *Agree*. But, say, *dates*, *Temperature in C^o* are on the Interval Scale.

Measurement Levels

- **Ratio Measurement** - Here, in addition to the Interval Scale (differences are meaningful), zero is not arbitrary (zero represents the absence of the characteristic measured), and the ratios are meaningful too.

Measurement Levels

- **Ratio Measurement** - Here, in addition to the Interval Scale (differences are meaningful), zero is not arbitrary (zero represents the absence of the characteristic measured), and the ratios are meaningful too. Say, *Weight*, *Height*, *Temperature in Kelvin* etc.

Measurement Levels

- **Ratio Measurement** - Here, in addition to the Interval Scale (differences are meaningful), zero is not arbitrary (zero represents the absence of the characteristic measured), and the ratios are meaningful too. Say, *Weight*, *Height*, *Temperature in Kelvin* etc. Say, the difference between 4KG and 2KG is the same as between 8KG and 6KG (equal differences are the same),

Measurement Levels

- **Ratio Measurement** - Here, in addition to the Interval Scale (differences are meaningful), zero is not arbitrary (zero represents the absence of the characteristic measured), and the ratios are meaningful too. Say, *Weight*, *Height*, *Temperature in Kelvin* etc. Say, the difference between 4KG and 2KG is the same as between 8KG and 6KG (equal differences are the same), and also 4KG is twice as heavy as 2KG.

Measurement Levels

- **Ratio Measurement** - Here, in addition to the Interval Scale (differences are meaningful), zero is not arbitrary (zero represents the absence of the characteristic measured), and the ratios are meaningful too. Say, *Weight*, *Height*, *Temperature in Kelvin* etc. Say, the difference between 4KG and 2KG is the same as between 8KG and 6KG (equal differences are the same), and also 4KG is twice as heavy as 2KG. 0KG means the ansence of the weight.

Measurement Levels

- **Ratio Measurement** - Here, in addition to the Interval Scale (differences are meaningful), zero is not arbitrary (zero represents the absence of the characteristic measured), and the ratios are meaningful too. Say, *Weight*, *Height*, *Temperature in Kelvin* etc. Say, the difference between 4KG and 2KG is the same as between 8KG and 6KG (equal differences are the same), and also 4KG is twice as heavy as 2KG. 0KG means the ansence of the weight. The same works also, say, for the *Temperature in Kelvin*.

Measurement Levels

- **Ratio Measurement** - Here, in addition to the Interval Scale (differences are meaningful), zero is not arbitrary (zero represents the absence of the characteristic measured), and the ratios are meaningful too. Say, *Weight*, *Height*, *Temperature in Kelvin* etc. Say, the difference between 4KG and 2KG is the same as between 8KG and 6KG (equal differences are the same), and also 4KG is twice as heavy as 2KG. 0KG means the ansence of the weight. The same works also, say, for the *Temperature in Kelvin*. But, say, for the *Temperature in Celcius*, 0 does not mean the *absence of the temperature*, and $40^o C$ is not twice as hot as $20^o C$.

Descriptive Statistics for a Univariate Data

Let me recall that we use the Descriptive Statistics at the beginning of our Statistical Study to examine, explore the dataset.

Descriptive Statistics for a Univariate Data

Let me recall that 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

Let me recall that 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

Let me recall that 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 a univariate *discrete* numerical or categorical data $x_1, x_2, ..., x_n$.

Frequency Tables

Here we assume that 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 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**:

```
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
```

Frequency Tables, Example

Now, consider the *iris* dataset in **R**:

```r
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Frequency Tables, Example

Now, consider the *iris* dataset in **R**:

```r
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Frequency Tables, Example, Cont'd

To get the *Species* Variable of the iris Dataset, we use

```
iris$Species
```

Frequency Tables, Example, Cont'd

To get the *Species* Variable of the iris Dataset, we use

```
iris$Species
```

And to calculate the Frequency of each of the Species, we use

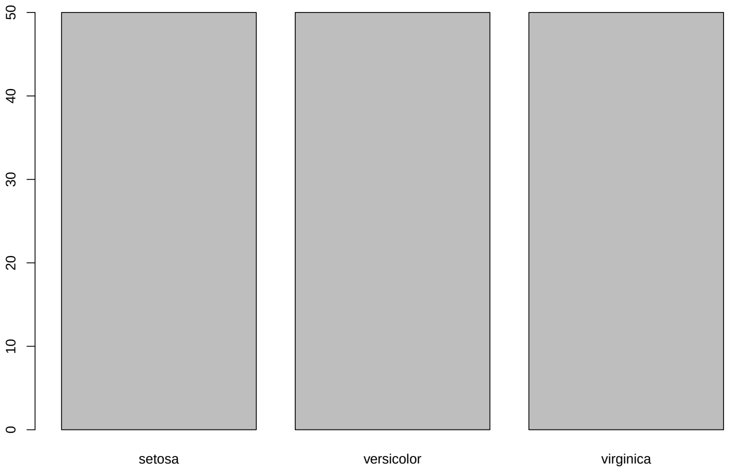
```
table(iris$Species)
```

```
##
##     setosa versicolor  virginica
##         50         50         50
```

LineGraph and BarPlot

Now, let us visualize our Frequency Table:

```
barplot(table(iris$Species))
```



LineGraph and Barplot

Another standard Dataset, *mtcars*, again about cars $\ddot\smile$:

```r
head(mtcars, 3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear c
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4
```

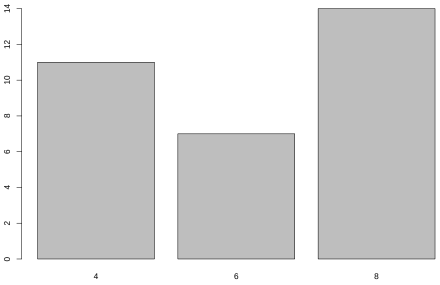
LineGraph and Barplot

Another standard Dataset, *mtcars*, again about cars :

```
head(mtcars, 3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear c
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4
```

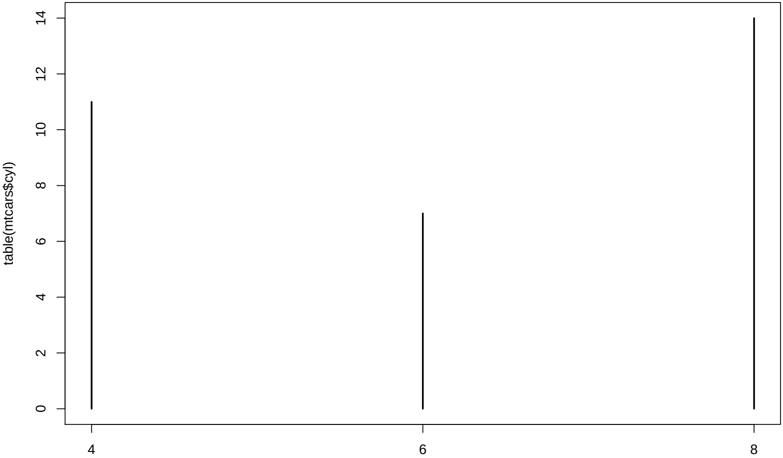
```
barplot(table(mtcars$cyl))
```



LineGraph and Barplot

Now, with the Line Graph:

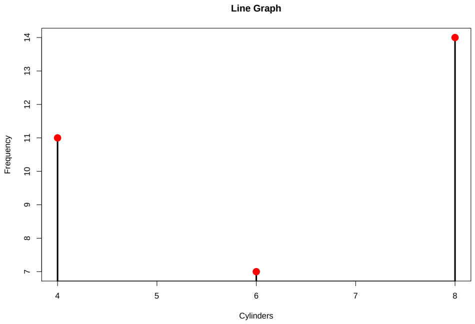
```
plot(table(mtcars$cyl))
```



LineGraph and Barplot

More sophisticated (titiz) version:

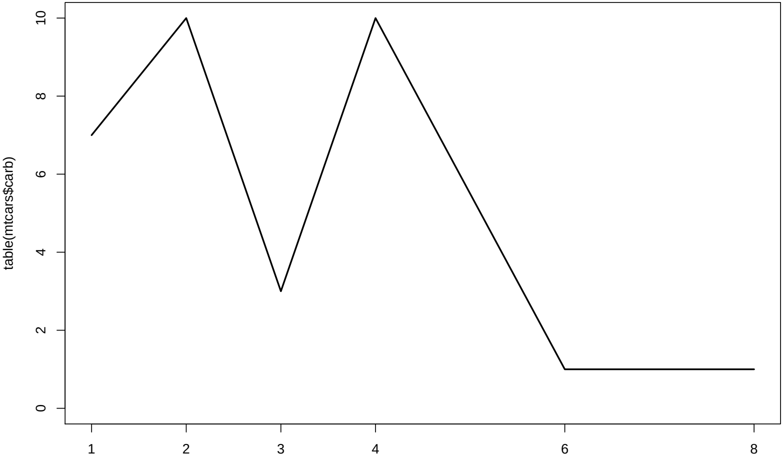
```
x <- mtcars$cyl; y <- as.data.frame(table(x))
a <- as.numeric(as.character(y$x)); b <- y$Freq
plot(a,b,type="h", lwd=3,  xlab = "Cylinders",
     ylab = "Frequency", main = "Line Graph")
points(a,b, pch=16, cex=2, col="red")
```



The Frequency Polygon

Again, same cars, but now the *carb* Variable Frequencies:

```r
plot(table(mtcars$carb), type = "l")
```



Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common, and is one of the main problems. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common, and is one of the main problems. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common, and is one of the main problems. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights.

From our Probability course, we know two complete characteristics of a Random Variable:

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common, and is one of the main problems. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights.

From our Probability course, we know two complete characteristics of a Random Variable: the **CDF and PD(M)F**.

Describing the Data Distribution

Assume we have a 1D numerical dataset x: $x_1, x_2, ..., x_n$. We assume that our dataset comes as a set of realizations of some Random Variable.

In Statistics, this is very common, and is one of the main problems. We assume that there is some RV behind our observations, we do not know the Distribution of that RV, but we have some observations from that Distribution. And our aim is to find (estimate) that Distribution.

Say, when we talk about the height distribution of persons between the ages 20-30, we assume that there is some unknown process that generates that heights.

From our Probability course, we know two complete characteristics of a Random Variable: the **CDF and PD(M)F**. So to describe our Data Distribution, we can try to describe the CDF and/or PD(M)F behind the Data.

Empirical CDF

First let's estimate the CDF. We will estimate CDF by the Empirical CDF:

Definition: The **Empirical Distribution Function, ECDF** or the **Cumulative Histogram** $ecdf(x)$ of our data $x_1, ..., x_n$ is defined by

$$ecdf(x) = \frac{\text{number of elements in our dataset} \le x}{\text{the total number of elements in our dataset}} =$$

$$= \frac{\text{number of elements in our dataset} \le x}{n}, \qquad \forall x \in \mathbb{R}.$$

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

- Analytical Part - on the board

To do the graphical part, we

- Sort our Dataset from the lowest to the largest values

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

- Analytical Part - on the board

To do the graphical part, we

- Sort our Dataset from the lowest to the largest values
- Plot the Data points on the OX axis

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

- Analytical Part - on the board

To do the graphical part, we

- Sort our Dataset from the lowest to the largest values
- Plot the Data points on the OX axis
- ECDF is 0 for values of x less that the smallest Datapoint, and is 1 for values of x bigger than the largest Datapoint

Example

Example: Construct the ECDF (analytically and graphically) of the following data:

$$-1, 4, 7, 5, 4$$

- Analytical Part - on the board

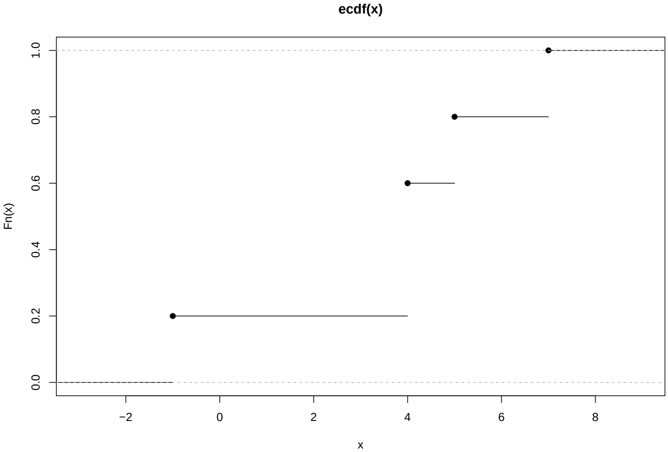
To do the graphical part, we

- Sort our Dataset from the lowest to the largest values
- Plot the Data points on the OX axis
- ECDF is 0 for values of x less that the smallest Datapoint, and is 1 for values of x bigger than the largest Datapoint
- For each Data point, calculate the Relative Frequency of that Datapoint (the number of times it occurs in our Dataset over the total number of Datapoints). At that Datapoint, do a Jump of the size of the Relative Frequency, and draw a horizontal line up to the next Datapoint

Example

Now, using **R**:

```
x <-c(-1, 4, 7, 5, 4)
f <- ecdf(x)
plot(f)
```



Note: It is easy to see that the ECDF satisfies all properties of a CDF.

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, \ldots, x_n$ comes from the Distribution with the CDF $F(x)$,

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, ..., x_n$ comes from the Distribution with the CDF $F(x)$,and if we will denote by $F_n(x)$ the ECDF constructed for $x_1, ..., x_n$, then

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, ..., x_n$ comes from the Distribution with the CDF $F(x)$,and if we will denote by $F_n(x)$ the ECDF constructed for $x_1, ..., x_n$, then

$$F_n(x) \to F(x) \quad \text{on} \quad \mathbb{R}.$$

Glivenko-Cantelli Theorem

How do we know that the ECDF is representing (estimating) the unknown CDF behind the Data good enough?

Well, this was proved by Glivenko and Cantelli: if our data $x_1, ..., x_n$ comes from the Distribution with the CDF $F(x)$,and if we will denote by $F_n(x)$ the ECDF constructed for $x_1, ..., x_n$, then

$$F_n(x) \to F(x) \quad \text{on} \quad \mathbb{R}.$$

Here we need to be more precise about in which sense the convergence holds.

Glivenko-Cantelli Theorem

In fact, the following Theorem Holds:

Theorem (Glivenko, Cantelli): If $X_1, ..., X_n$ are r.v.s from the Distribution with the CDF $F(x)$, and $F_n(x)$ is the ECDF constructed by using $X_1, ..., X_n$, then

$$\sup_x |F_n(x) - F(x)| \to 0 \qquad a.s.$$

Glivenko-Cantelli Theorem

In fact, the following Theorem Holds:

Theorem (Glivenko, Cantelli): If $X_1, ..., X_n$ are r.v.s from the Distribution with the CDF $F(x)$, and $F_n(x)$ is the ECDF constructed by using $X_1, ..., X_n$, then

$$\sup_x |F_n(x) - F(x)| \to 0 \qquad a.s.$$

This Theorem says that if you will have enough datapoints from a Distribution, you can approximate the unknown CDF of your Distribution pretty well by using the ECDF.

Estimation of the CDF through ECDF

Let us check this theorem using **R**:

```
plot(pnorm, lwd = 3, col = 'red', xlim = c(-3,3),
     ylim = c(0,1), ylab = "ecdf and CDF")
n <- 30 ; x <- rnorm(n) #Taking a sample of size n from N(0,1)
f <- ecdf(x) #f will be the ECDF of our data x
par(new = TRUE) #this is to keep the previous graph
plot(f, xlim = c(-3,3), ylim = c(0,1), ylab = "ecdf and CDF")
```