AUA CS 108, Statistics, Fall 2019

Lecture 02

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

28 Aug 2019

Intro To Statistics

Contents

- Some important Notions and Definitions
- Stages of Doing a Statistical Analysis
- Different Types of Variables

Last Lecture ReCap

- *Question:* What is Statistics?

Last Lecture ReCap

- *Question:* What is Statistics?
- *Question:* What is Descriptive Statistics? Why we need it?

Last Lecture ReCap

- *Question:* What is Statistics?
- *Question:* What is Descriptive Statistics? Why we need it?
- Are **PSSs on Mondays** convenient for you?

Quiz Time

Quiz Time is yet to Come $\ddot\smile$

Some Important Notions and Definitions

- **Population** is the totality of all elements under interest

Some Important Notions and Definitions

- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

Some Important Notions and Definitions

- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Statistics, roughly, we use the Sample to get information about the Population.

Some Important Notions and Definitions

- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Statistics, roughly, we use the Sample to get information about the Population.

- **Sampling** is the process of choosing a Sample

Some Important Notions and Definitions

- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Statistics, roughly, we use the Sample to get information about the Population.

- **Sampling** is the process of choosing a Sample
- **Observation** is the Data (information) collected from one element in the Sample

Some Important Notions and Definitions

- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Statistics, roughly, we use the Sample to get information about the Population.

- **Sampling** is the process of choosing a Sample
- **Observation** is the Data (information) collected from one element in the Sample
- **Variable** (or the **Feature**) is a characteristic whose value may change from one element to other one in population.

Some Important Notions and Definitions

- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Statistics, roughly, we use the Sample to get information about the Population.

- **Sampling** is the process of choosing a Sample
- **Observation** is the Data (information) collected from one element in the Sample
- **Variable** (or the **Feature**) is a characteristic whose value may change from one element to other one in population.
- **Parameter** is a numerical (1D or n-D) characteristic of the *Population*

Some Important Notions and Definitions

- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Statistics, roughly, we use the Sample to get information about the Population.

- **Sampling** is the process of choosing a Sample
- **Observation** is the Data (information) collected from one element in the Sample
- **Variable** (or the **Feature**) is a characteristic whose value may change from one element to other one in population.
- **Parameter** is a numerical (1D or n-D) characteristic of the *Population*
- **Statistics** is a numerical characteristic of the *Sample*

Example:

Here is one of the standard Datasets in **R**:

```
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Example:

Here is one of the standard Datasets in **R**:

```r
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

- Which are the **Variables** ?

Example:

Here is one of the standard Datasets in **R**:

```r
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

- Which are the **Variables** ?
- Give two **Observations**.

Example

Example: Say, we want to get information about the ratio of female students in AUA. We conduct an experiment: calculate the ratio of female students in our class.

Example

Example: Say, we want to get information about the ratio of female students in AUA. We conduct an experiment: calculate the ratio of female students in our class.

Here,

- the **Population** is

Example

Example: Say, we want to get information about the ratio of female students in AUA. We conduct an experiment: calculate the ratio of female students in our class.

Here,

- the **Population** is
- the **Sample** is

Example

Example: Say, we want to get information about the ratio of female students in AUA. We conduct an experiment: calculate the ratio of female students in our class.

Here,

- the **Population** is
- the **Sample** is
- the **Parameter** is

Example

Example: Say, we want to get information about the ratio of female students in AUA. We conduct an experiment: calculate the ratio of female students in our class.

Here,

- the **Population** is
- the **Sample** is
- the **Parameter** is
- the **Statistics** is

Example

Example: Say, we want to get information about the ratio of female students in AUA. We conduct an experiment: calculate the ratio of female students in our class.

Here,

- the **Population** is
- the **Sample** is
- the **Parameter** is
- the **Statistics** is
- an **Obsevation** is

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 **Obsevation** ?

Example

Example: We are doing some measurement (say, calculating the speed of a light $\ddot\smile$).

Example

Example: We are doing some measurement (say, calculating the speed of a light $\ddot\smile$). Our measurement device is not ideal, so we will get some errors every time we do our measurement.

Example

Example: We are doing some measurement (say, calculating the speed of a light $\ddot\smile$). Our measurement device is not ideal, so we will get some errors every time we do our measurement. We assume that the measurement results are Normally Distributed: if X is the measurement result (a r.v.!), then $X \sim \mathcal{N}(\mu, \sigma^2)$.

Example

Example: We are doing some measurement (say, calculating the speed of a light $\ddot\smile$). Our measurement device is not ideal, so we will get some errors every time we do our measurement. We assume that the measurement results are Normally Distributed: if X is the measurement result (a r.v.!), then $X \sim \mathcal{N}(\mu, \sigma^2)$.

Can you describe

- the **Population**

Example

Example: We are doing some measurement (say, calculating the speed of a light $\ddot\smile$). Our measurement device is not ideal, so we will get some errors every time we do our measurement. We assume that the measurement results are Normally Distributed: if X is the measurement result (a r.v.!), then $X \sim \mathcal{N}(\mu, \sigma^2)$.

Can you describe

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

Example

Example: We are doing some measurement (say, calculating the speed of a light $\ddot{\smile}$). Our measurement device is not ideal, so we will get some errors every time we do our measurement. We assume that the measurement results are Normally Distributed: if X is the measurement result (a r.v.!), then $X \sim \mathcal{N}(\mu, \sigma^2)$.

Can you describe

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

Example

Example: We are doing some measurement (say, calculating the speed of a light $\ddot\smile$). Our measurement device is not ideal, so we will get some errors every time we do our measurement. We assume that the measurement results are Normally Distributed: if X is the measurement result (a r.v.!), then $X \sim \mathcal{N}(\mu, \sigma^2)$.

Can you describe

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

Example

Example: We are doing some measurement (say, calculating the speed of a light $\ddot\smile$). Our measurement device is not ideal, so we will get some errors every time we do our measurement. We assume that the measurement results are Normally Distributed: if X is the measurement result (a r.v.!), then $X \sim \mathcal{N}(\mu, \sigma^2)$.

Can you describe

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

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 $\ddot\smile$). 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.

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?

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 AUA.

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 AUA.

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 AUA.

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?