AUA CS108, Statistics, Fall 2020

Lecture 02

Michael Poghosyan

28 Aug 2020

Contents

Some Problems that Statistics can consider

One real-life Problem: just few days ago a friend of mine called to ask this question.

Some Problems that Statistics can consider

One real-life Problem: just few days ago a friend of mine called to ask this question.

The problem is the following: they have a large Dataset of Observations (say, Excel sheet with information like the person's name, surname, age, wage, size of a loan, whether returned that loan on time or not etc).

Some Problems that Statistics can consider

One real-life Problem: just few days ago a friend of mine called to ask this question.

The problem is the following: they have a large Dataset of Observations (say, Excel sheet with information like the person's name, surname, age, wage, size of a loan, whether returned that loan on time or not etc). The problem is that they want to check if the Data is correct or not.

Some Problems that Statistics can consider

One real-life Problem: just few days ago a friend of mine called to ask this question.

The problem is the following: they have a large Dataset of Observations (say, Excel sheet with information like the person's name, surname, age, wage, size of a loan, whether returned that loan on time or not etc). The problem is that they want to check if the Data is correct or not. Because the number of observations was very large, the question was: how many observations is enough to check for correctness to be sure with 95% that the data is correct?

Some Problems that Statistics can consider

One real-life Problem: just few days ago a friend of mine called to ask this question.

The problem is the following: they have a large Dataset of Observations (say, Excel sheet with information like the person's name, surname, age, wage, size of a loan, whether returned that loan on time or not etc). The problem is that they want to check if the Data is correct or not. Because the number of observations was very large, the question was: how many observations is enough to check for correctness to be sure with 95% that the data is correct?

Of course, the question/statement was not correct.

Some Problems that Statistics can consider

One real-life Problem: just few days ago a friend of mine called to ask this question.

The problem is the following: they have a large Dataset of Observations (say, Excel sheet with information like the person's name, surname, age, wage, size of a loan, whether returned that loan on time or not etc). The problem is that they want to check if the Data is correct or not. Because the number of observations was very large, the question was: how many observations is enough to check for correctness to be sure with 95% that the data is correct?

Of course, the question/statement was not correct. If, say, half of the data is not correct, then, even if you will check the whole Dataset for correctness, you cannot be sure with 95% that the data is correct.

Some Problems that Statistics can consider

But what can we do?

Some Problems that Statistics can consider

But what can we do?

Using Statistics, methods we will learn, we can

Some Problems that Statistics can consider

But what can we do?

Using Statistics, methods we will learn, we can

- Estimate the Proportion of correct data in the Dataset using just part of the Data, using a **Sample** from data;

Some Problems that Statistics can consider

But what can we do?

Using Statistics, methods we will learn, we can

- Estimate the Proportion of correct data in the Dataset using just part of the Data, using a **Sample** from data;
- moreover, we can Estimate the number of Observations we need to take in the Sample (Sample Size) to be sure with 95% that the real proportion (of correct data) is within given small neighborhood of our Estimate;

Some Problems that Statistics can consider

But what can we do?

Using Statistics, methods we will learn, we can

- Estimate the Proportion of correct data in the Dataset using just part of the Data, using a **Sample** from data;
- moreover, we can Estimate the number of Observations we need to take in the Sample (Sample Size) to be sure with 95% that the real proportion (of correct data) is within given small neighborhood of our Estimate;
- also we need to suggest how to choose that Sample (we need a Random Sampling)

Course Structure: Topics at a glance

The structure of our course will be the following:

- Intro + Descriptive Statistics

Course Structure: Topics at a glance

The structure of our course will be the following:

- Intro + Descriptive Statistics
- (Very) Quick reminder on RVs, Convergence Types for RVs, and our good old LLN and CLT:

The rest will use these topics intensively.

Course Structure: Topics at a glance

The structure of our course will be the following:

- Intro + Descriptive Statistics
- (Very) Quick reminder on RVs, Convergence Types for RVs, and our good old LLN and CLT:

The rest will use these topics intensively.

- Models, Statistical Inference and Learning:

Here we will talk mainly about the Parametric Statistics.

Course Structure: Topics at a glance

Then we will run over three main problems of the Parametric Statistics:

- Parameter Point Estimates

Course Structure: Topics at a glance

Then we will run over three main problems of the Parametric Statistics:

- Parameter Point Estimates
- Confidence Intervals

Course Structure: Topics at a glance

Then we will run over three main problems of the Parametric Statistics:

- Parameter Point Estimates
- Confidence Intervals
- Hypothesis Testing

Course Structure: Topics at a glance

Then we will run over three main problems of the Parametric Statistics:

- Parameter Point Estimates
- Confidence Intervals
- Hypothesis Testing

Then we will talk a little bit about the *Bayesian Approach*, basics of

- Bayesian Estimation

Course Structure: Topics at a glance

Then we will run over three main problems of the Parametric Statistics:

- Parameter Point Estimates
- Confidence Intervals
- Hypothesis Testing

Then we will talk a little bit about the *Bayesian Approach*, basics of

- Bayesian Estimation

Next we will focus on the simplest Statistical Model for the relationship between different Variables: we will learn

- Linear Regression

Course Structure: Topics at a glance

Then we will run over three main problems of the Parametric Statistics:

- Parameter Point Estimates
- Confidence Intervals
- Hypothesis Testing

Then we will talk a little bit about the *Bayesian Approach*, basics of

- Bayesian Estimation

Next we will focus on the simplest Statistical Model for the relationship between different Variables: we will learn

- Linear Regression

And at the end of the course we will return back to Testing and cover:

- Goodness of fit tests

Stages of the Statistical Analysis, and Data Types

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 Data

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

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

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

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

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

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

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

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

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

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

We will mainly talk about the 2nd and 3rd stages.

Stages of Doing a Statistical Analysis

Important Stages of the Statistical Analysis are:

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

We will mainly talk about the 2nd and 3rd stages. But first I want to give some Notions and Definitions we will use later.

Some Important Notions and Definitions

- **Data** are a collection of information about some objects or subjects under interest

Some Important Notions and Definitions

- **Data** are a collection of information about some objects or subjects under interest
- **Population** is the totality of all elements under interest

Some Important Notions and Definitions

- **Data** are a collection of information about some objects or subjects under interest
- **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

- **Data** are a collection of information about some objects or subjects under interest
- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

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

Some Important Notions and Definitions

- **Data** are a collection of information about some objects or subjects under interest
- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Inferential 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

- **Data** are a collection of information about some objects or subjects under interest
- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Inferential 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

- **Data** are a collection of information about some objects or subjects under interest
- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Inferential 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 a **Feature**) is a characteristic whose value may change from one element to other one in population.

Some Important Notions and Definitions

- **Data** are a collection of information about some objects or subjects under interest
- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Inferential 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 a **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

- **Data** are a collection of information about some objects or subjects under interest
- **Population** is the totality of all elements under interest
- **Sample** is a subset of a Population, that will be studied

In Inferential 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 a **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** (the first several rows):

```
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** (the first several rows):

```
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** (the first several rows):

```
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 calculate the proportion of female students in AUA. We conduct an experiment: calculate the proportion of female students in our class.

Example

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

Here,

- the **Population** is

Example

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

Here,

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

Example

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

Here,

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

Example

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

Here,

- the **Population** is
- the **Sample** is
- the **Parameter** is
- the **Statistics** is
- an **Observation** 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 **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 $\ddot{\smile}$).

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 ☺). 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 ☺). 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 ☺). 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 ☺). 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 ☺). 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 ☺). 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 ☺). 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. Why?

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. 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. 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 AUA.

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 AUA.

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 AUA.

Is that Sample representative? Why?