ASDS Statistics, YSU, Fall 2020

Lecture 01

Michael Poghosyan

15 Sep 2020

Welcome to the ASDS Statistics Course

Welcome to the ASDS Statistics Course

And Happy New ~~Year~~ Semester ! ⌣̈

Contents

- Syllabus highlights
- Intro to the Course
- Glance at a Course Structure
- Some important Notions and Definitions
- Stages of Doing a Statistical Analysis
- Different Types of Variables

Syllabus Highlights

- Course name: **Applied Statistics with R**

Syllabus Highlights

- Course name: **Applied Statistics with R**
- No. of Credits: **3**

Syllabus Highlights

- Course name: **Applied Statistics with R**
- No. of Credits: **3**
- Instructor: **MP**

Syllabus Highlights

- Course name: **Applied Statistics with R**
- No. of Credits: **3**
- Instructor: **MP**
- Instructor's OH: By an Appointment

Syllabus Highlights

- Course name: **Applied Statistics with R**
- No. of Credits: **3**
- Instructor: **MP**
- Instructor's OH: By an Appointment
- Teaching Associates: ??

Course Materials

- Discussions and Info Platform: **Slack**

Course Materials

- Discussions and Info Platform: **Slack**
- Syllabus: will be on GD

Course Materials

- Discussions and Info Platform: **Slack**
- Syllabus: will be on GD
- Textbooks: will be on GD

Course Materials

- Discussions and Info Platform: **Slack**
- Syllabus: will be on GD
- Textbooks: will be on GD
- Software: **R** and **R Studio** (freeware)

Course Materials

- Discussions and Info Platform: **Slack**
- Syllabus: will be on GD
- Textbooks: will be on GD
- Software: **R** and **R Studio** (freeware)
- R Textbooks: will be on GD

Course Materials

- Discussions and Info Platform: **Slack**
- Syllabus: will be on GD
- Textbooks: will be on GD
- Software: **R** and **R Studio** (freeware)
- R Textbooks: will be on GD

Syllabus Highlights, Cont'd

- Exams: **2 Midterm** and a **Final Exam**

Syllabus Highlights, Cont'd

- Exams: **2 Midterm** and a **Final Exam**
- Homework: we will have them (2 points in the total)

Syllabus Highlights, Cont'd

- Exams: **2 Midterm** and a **Final Exam**
- Homework: we will have them (2 points in the total)
- No late HWs (except some veeery special cases)

Syllabus Highlights, Cont'd

- Exams: **2 Midterm** and a **Final Exam**
- Homework: we will have them (2 points in the total)
- No late HWs (except some veeery special cases)
- Advice: Always ask your questions during lectures, ask your questions during OHs, solve HWs by yourself!

Syllabus Highlights, Cont'd

- Exams: **2 Midterm** and a **Final Exam**
- Homework: we will have them (2 points in the total)
- No late HWs (except some veeery special cases)
- Advice: Always ask your questions during lectures, ask your questions during OHs, solve HWs by yourself!
- Advice: Run over the Probability Topics, especially, about RVs and Distributions

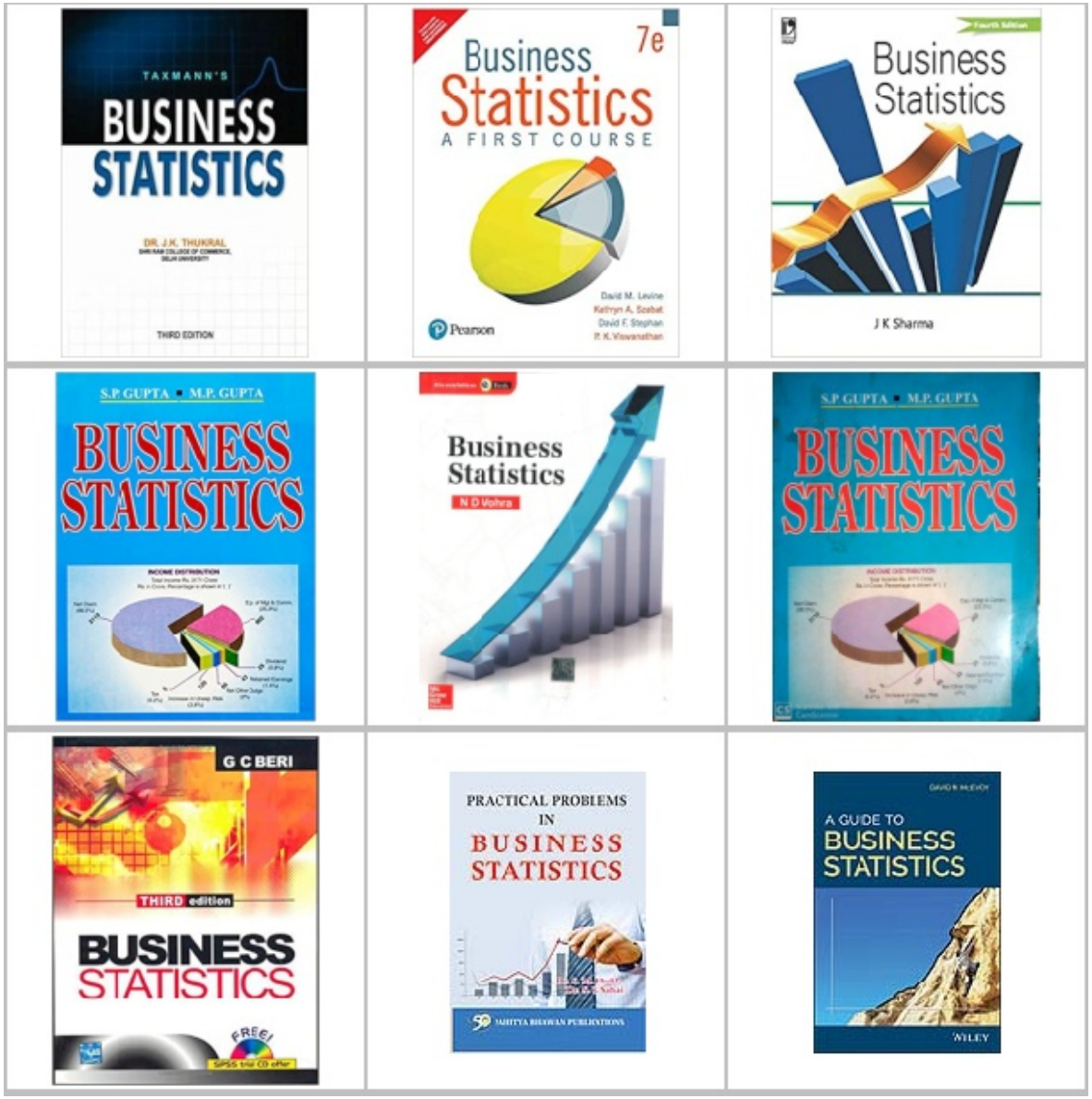
QA Session

Questions?

Where we use Statistics

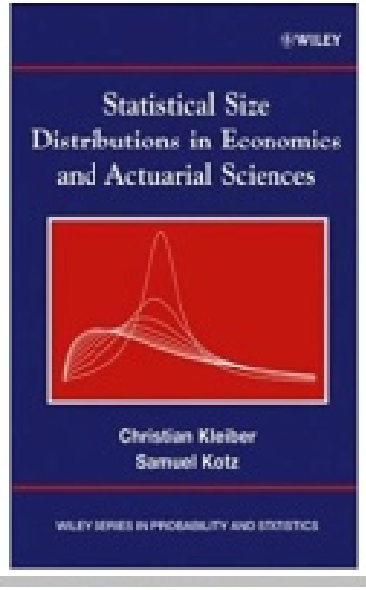
Where we use Statistics

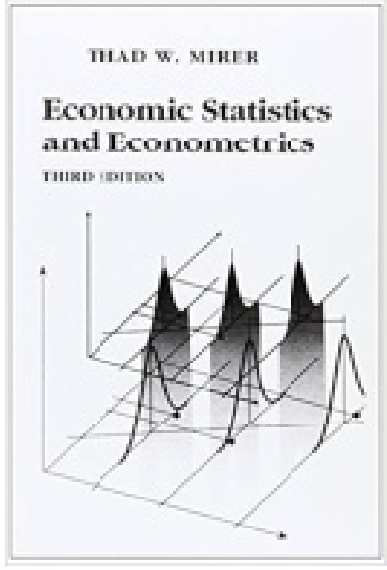
Business

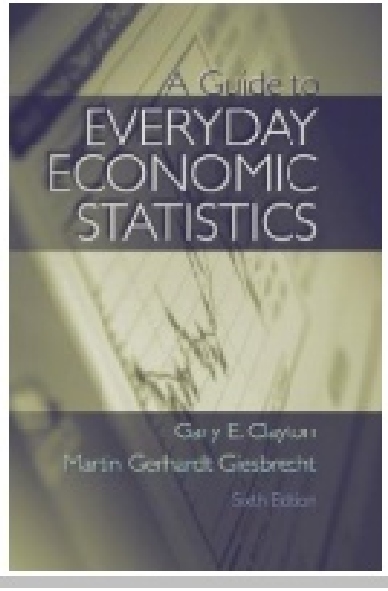


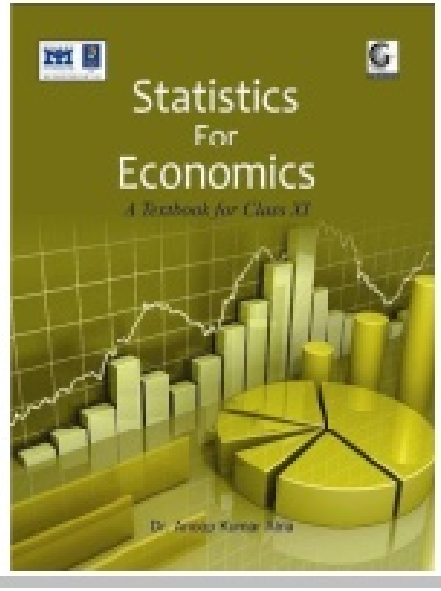
Where we use Statistics

Economics









Where we use Statistics

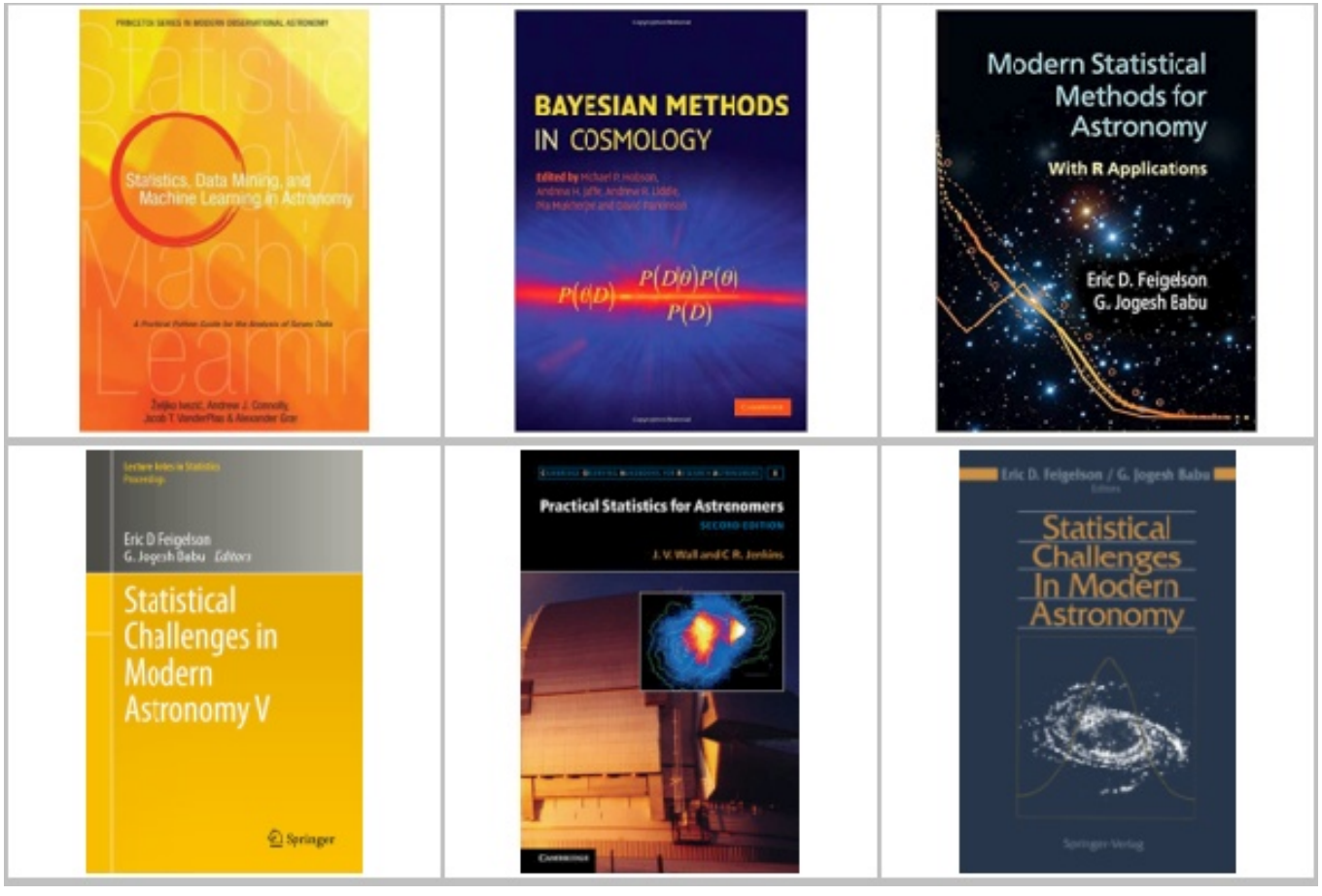
Agriculture

Where we use Statistics

Finance

Where we use Statistics

Astronomy



Where we use Statistics

Biology

Where we use Statistics

BioStatistics

Where we use Statistics

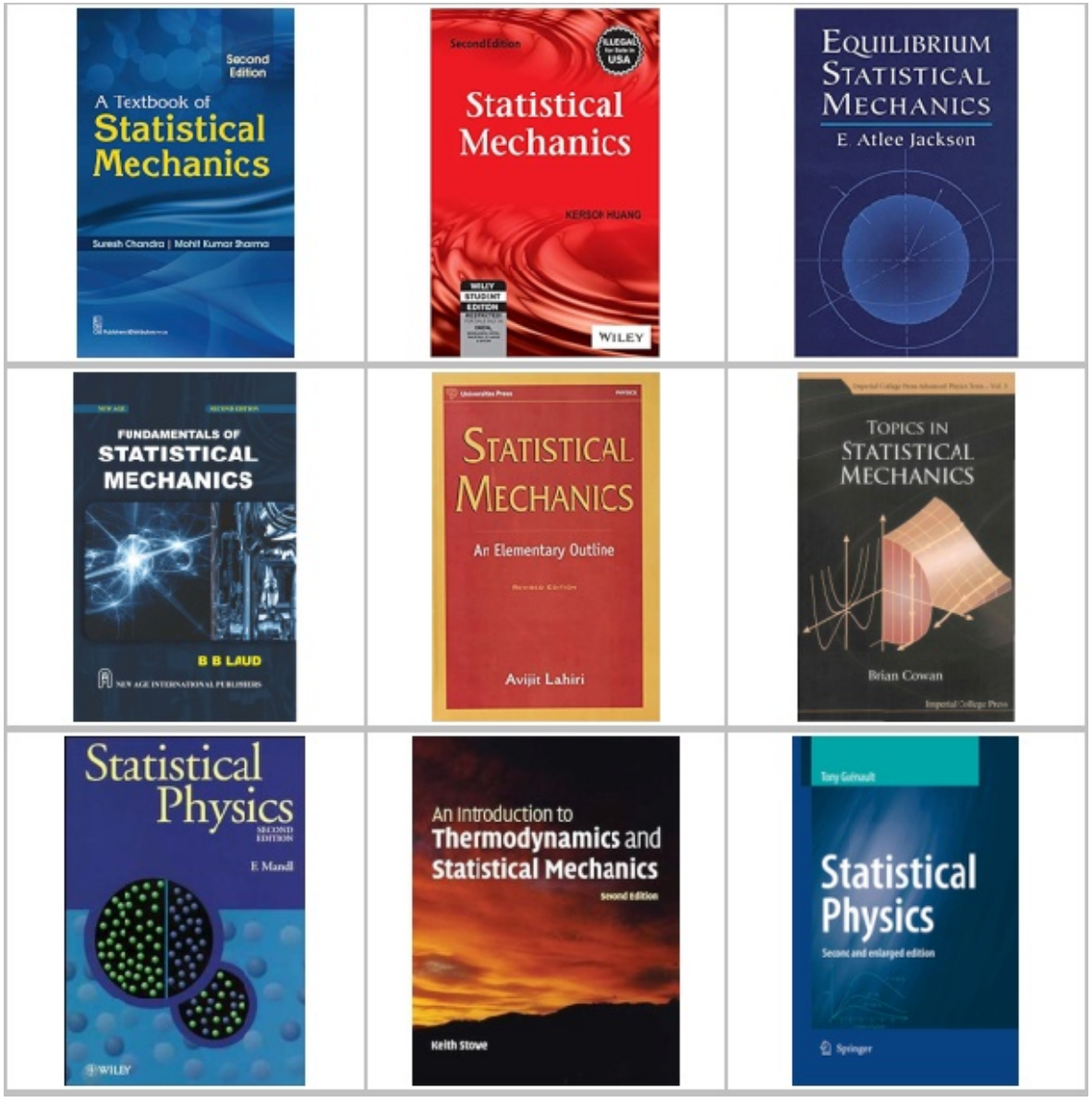
Psychology

Where we use Statistics

Medicine

Where we use Statistics

Physics and Mechanics



Where we use Statistics

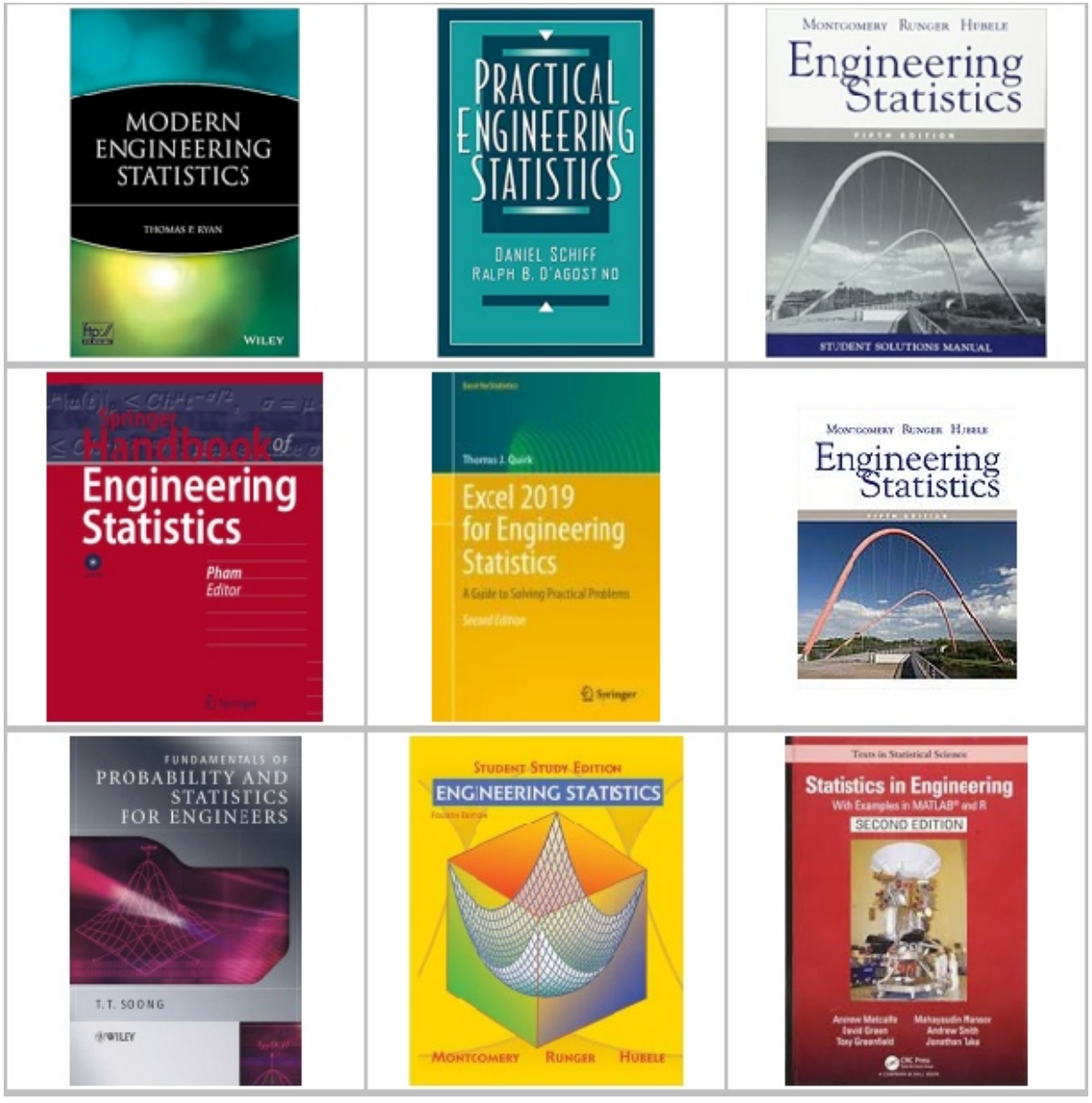
Marketing

Where we use Statistics

Language Study

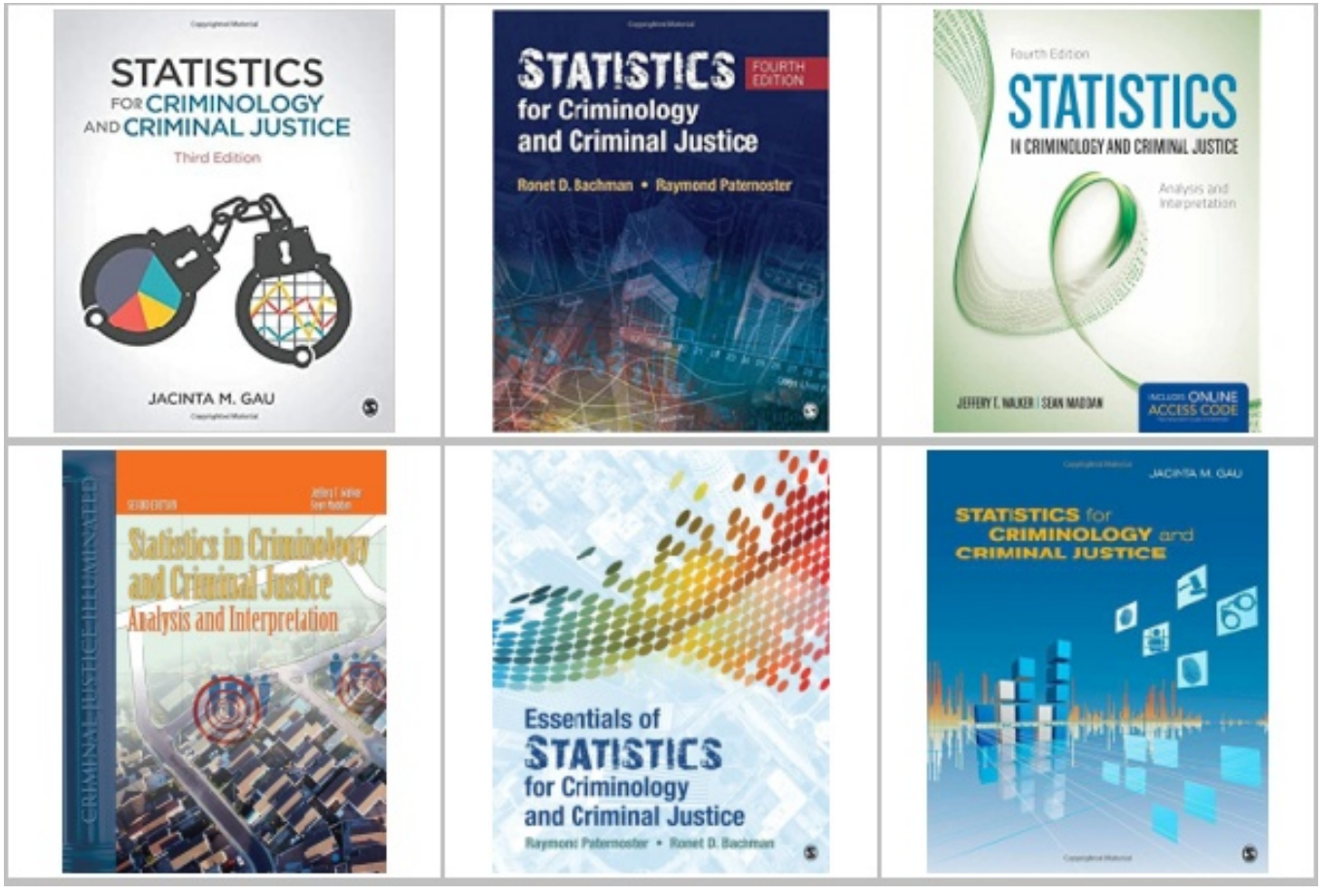
Where we use Statistics

Engineering



Where we use Statistics

Criminology

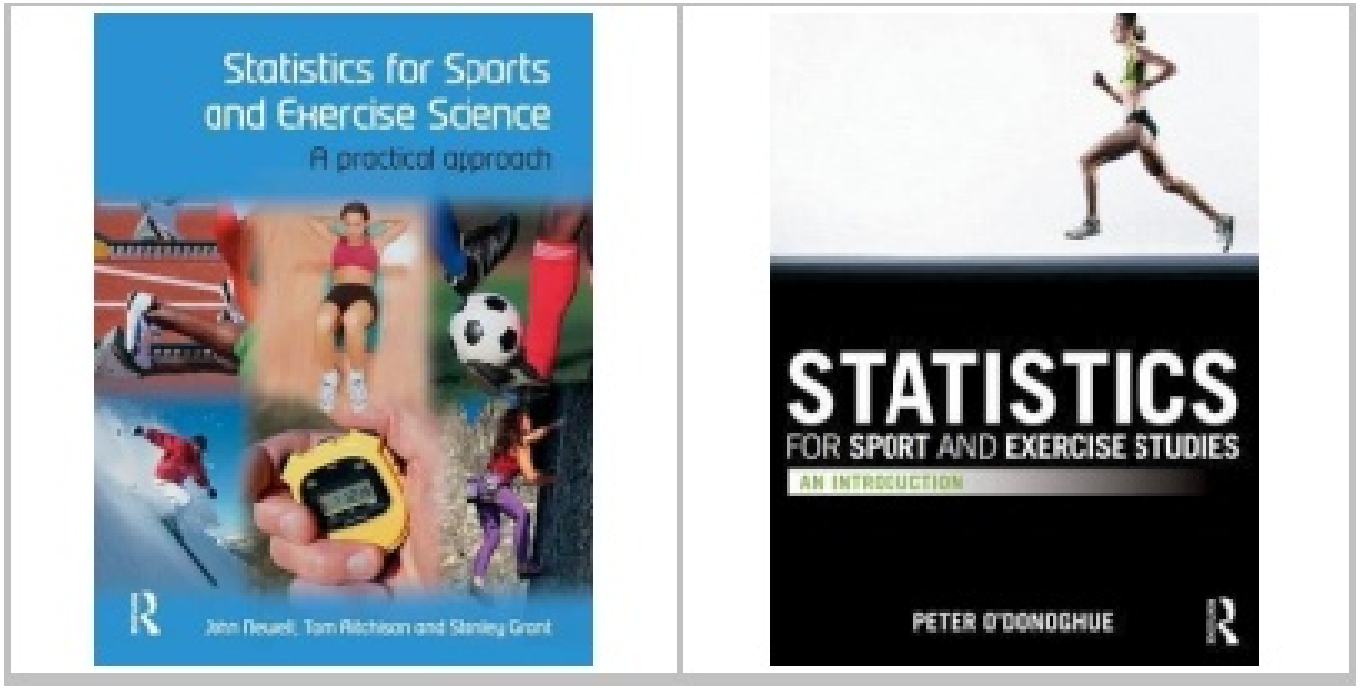


Where we use Statistics

Quality Control

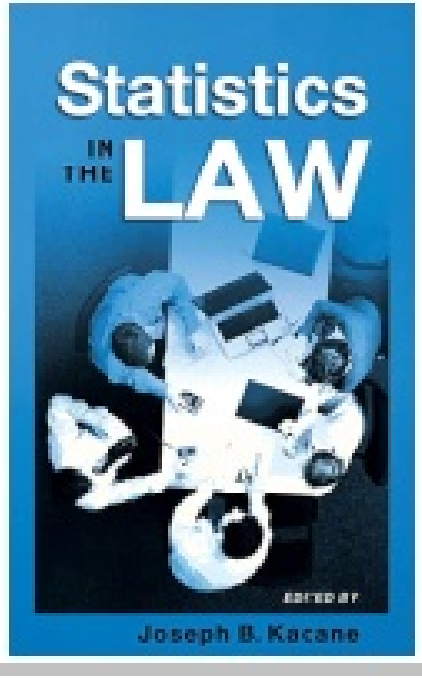
Where we use Statistics

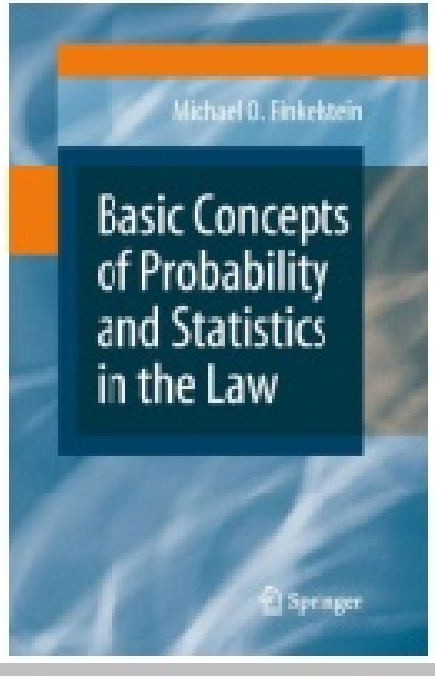
Sport



Where we use Statistics

Law





Where we use Statistics

Food

Where we use Statistics

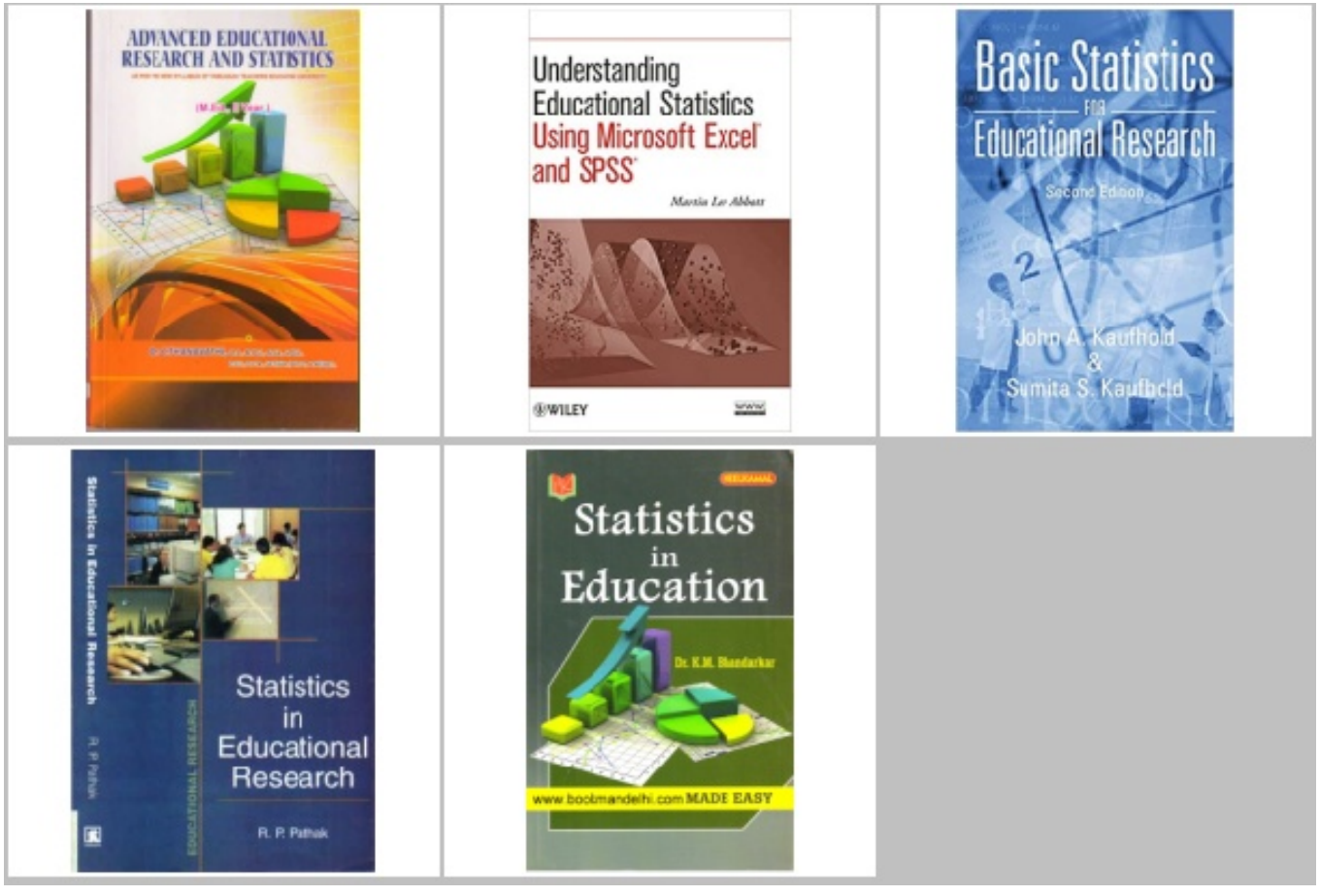
Genetics

Where we use Statistics

Chemistry

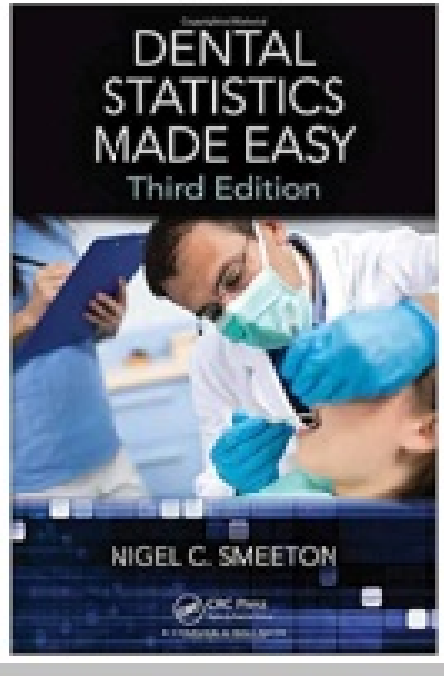
Where we use Statistics

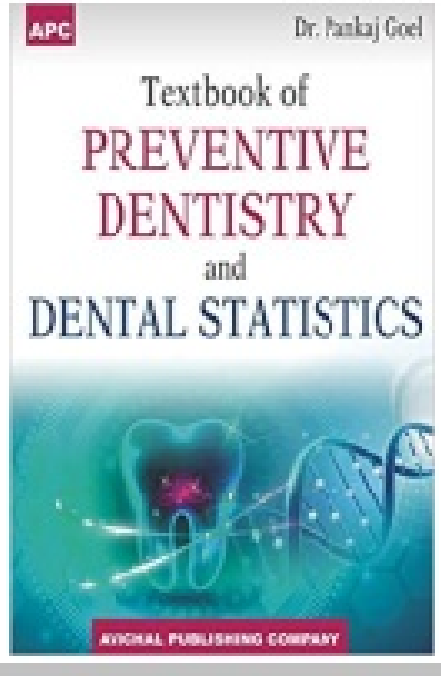
Education



Where we use Statistics

Dentistry





Where we use Statistics

Other. . .

Where we use Statistics

ML, Statistical Learning

Where we use Statistics

. . . and so on . . .

About Statistics

- What is Statistics?

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data. In a little bit detailed form,

Statistics is the Art and Science of ***Collecting, Describing and Analyzing Data****, getting insight from Data.*

About Statistics

- What is the difference between Statistics and Probability?

About Statistics

- What is the difference between Statistics and Probability?

In some sense, Statistics and Probability are inverse to each other:

About Statistics

- What is the difference between Statistics and Probability?

In some sense, Statistics and Probability are inverse to each other:

- in Probability Theory, we assume the Reality, the Truth, the Generating Process is given, and we get information about possible outcomes, possible observations.

About Statistics

- What is the difference between Statistics and Probability?

In some sense, Statistics and Probability are inverse to each other:

- in Probability Theory, we assume the Reality, the Truth, the Generating Process is given, and we get information about possible outcomes, possible observations. Say,

We assume that the daily number of customers who are ordering Pepperoni Pizza at a particular pizzeria is following the $Pois(15.2)$ Distribution (*the Reality, the Generating Process*), and we are interested in the Probability that the daily number will exceed 17 (*Probability of a possible outcome*)

About Statistics

- in Statistics, we have Observations, we have Data, Outcomes, we want to learn about the Reality, about the Truth, about the Generating Process (from which that Data is obtained).

About Statistics

- in Statistics, we have Observations, we have Data, Outcomes, we want to learn about the Reality, about the Truth, about the Generating Process (from which that Data is obtained). Say,

We have Data about daily number of customers ordering Pepperoni Pizza at that pizzeria for some days:

$$10,\ 15,\ 14,\ 20,\ 7,\ 12,\ 10,$$

and we want to find out the Distribution of the r.v.

$X =$ the daily number of customers ordering PP at that pizzeria.

About Statistics

It is easy to see that this problem is more challenging than its Probabilistic counterpart, its “inverse” one -

About Statistics

It is easy to see that this problem is more challenging than its Probabilistic counterpart, its “inverse” one - because we want to learn about the Reality, the Truth from a finite number of observations.

About Statistics

It is easy to see that this problem is more challenging than its Probabilistic counterpart, its "inverse" one - because we want to learn about the Reality, the Truth from a finite number of observations. And, moreover, if the Probabilistic counterpart has a unique correct answer (we just need to calculate $\mathbb{P}(X > 17)$, where X is the daily number of PP-hungry customers, $X \sim Pois(15.2)$),

About Statistics

It is easy to see that this problem is more challenging than its Probabilistic counterpart, its "inverse" one - because we want to learn about the Reality, the Truth from a finite number of observations. And, moreover, if the Probabilistic counterpart has a unique correct answer (we just need to calculate $\mathbb{P}(X > 17)$, where X is the daily number of PP-hungry customers, $X \sim Pois(15.2)$), its Statistical counterpart doesn't have a unique correct answer!

About Statistics

It is easy to see that this problem is more challenging than its Probabilistic counterpart, its "inverse" one - because we want to learn about the Reality, the Truth from a finite number of observations. And, moreover, if the Probabilistic counterpart has a unique correct answer (we just need to calculate $\mathbb{P}(X > 17)$, where X is the daily number of PP-hungry customers, $X \sim Pois(15.2)$), its Statistical counterpart doesn't have a unique correct answer! We cannot find out the Exact Distribution, we can just guess it, we can only **Estimate**.

About Statistics

- What is the difference between Statistical and Mathematical thinking?

About Statistics

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms).

About Statistics

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases!

About Statistics

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are

About Statistics

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are $x = \pm 2$.

About Statistics

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are $x = \pm 2$.

In Statistics (Inferential Statistics), we use data to get an insight about the unknown process behind the generation of the data. Usually, our data is finite, and this is not giving a chance to get a complete and 100 percent information about that unknown process.

About Statistics

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are $x = \pm 2$.

In Statistics (Inferential Statistics), we use data to get an insight about the unknown process behind the generation of the data. Usually, our data is finite, and this is not giving a chance to get a complete and 100 percent information about that unknown process. So, in Statsitsics, we are doing some *assumptions*, and from that "inferring", "guessing", ***estimating*** the unknown process, its parameters, and sometimes give the level of our confidence.

About Statistics

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are $x = \pm 2$.

In Statistics (Inferential Statistics), we use data to get an insight about the unknown process behind the generation of the data. Usually, our data is finite, and this is not giving a chance to get a complete and 100 percent information about that unknown process. So, in Statsitsics, we are doing some *assumptions*, and from that "inferring", "guessing", ***estimating*** the unknown process, its parameters, and sometimes give the level of our confidence. Say, if I am tossing a coin 10 times and get 8 Hs and 2 Ts, is it a sign that the coin is not fair?

About Statistics

- What is the difference between Data Scienece and Statistics?

About Statistics

- What is the difference between Data Scienece and Statistics?

Well, maybe we will talk a little bit about this at the end of the course.

- Why I need to learn Statistics?

About Statistics

- What is the difference between Data Scienece and Statistics?

Well, maybe we will talk a little bit about this at the end of the course.

- Why I need to learn Statistics?

Simple - to pass this course ¨‿

About Statistics

- What is the difference between Data Scienece and Statistics?

Well, maybe we will talk a little bit about this at the end of the course.

- Why I need to learn Statistics?

Simple - to pass this course ⌣̈ Or, it is very important if you want to become a Statistician (one of the fastest growing and Best Jobs in USA), learn DS/ML, become a Biostatistician, learn Econometrics, make Investments and play in Financial Markets, and many-many more, see, for example, the Wiki page.

About Statistics

- What is the difference between Data Scienece and Statistics?

Well, maybe we will talk a little bit about this at the end of the course.

- Why I need to learn Statistics?

Simple - to pass this course ⌣̈ Or, it is very important if you want to become a Statistician (one of the fastest growing and Best Jobs in USA), learn DS/ML, become a Biostatistician, learn Econometrics, make Investments and play in Financial Markets, and many-many more, see, for example, the Wiki page.

And finally, to understand the everyday usage of Statistical language, graphs and estimates, say, about polls and salaries ⌣̈

Some Problems that Statistics can consider

- Estimate the proportion of adults in Armenia that are computer-literate;

Some Problems that Statistics can consider

- Estimate the proportion of adults in Armenia that are computer-literate;
- Estimate the number/proportion of defective production in a manufacturing plant;

Some Problems that Statistics can consider

- Estimate the proportion of adults in Armenia that are computer-literate;
- Estimate the number/proportion of defective production in a manufacturing plant;
- Show that there is a strong correlation between education and income;

Some Problems that Statistics can consider

- Estimate the proportion of adults in Armenia that are computer-literate;
- Estimate the number/proportion of defective production in a manufacturing plant;
- Show that there is a strong correlation between education and income;
- Determine whether it is true that, in average, women earn less than men;

Some Problems that Statistics can consider

- Find the relationship between the High School Math average grade and the average of the University Calc 1 and Calc 2 courses;

Some Problems that Statistics can consider

- Find the relationship between the High School Math average grade and the average of the University Calc 1 and Calc 2 courses;

Note: Clearly, there is **no** exact formula relating the average of Calc1 and Calc2 grades to HS Math average grade.

Some Problems that Statistics can consider

- Find the relationship between the High School Math average grade and the average of the University Calc 1 and Calc 2 courses;

Note: Clearly, there is **no** exact formula relating the average of Calc1 and Calc2 grades to HS Math average grade. In statistics, we do not *find* exactly, but try to give a reasonable *Estimate*, even we assess how good is our Estimate;

Some Problems that Statistics can consider

- Find the relationship between the High School Math average grade and the average of the University Calc 1 and Calc 2 courses;

Note: Clearly, there is **no** exact formula relating the average of Calc1 and Calc2 grades to HS Math average grade. In statistics, we do not *find* exactly, but try to give a reasonable *Estimate*, even we assess how good is our Estimate;

- Calculate the percentage of body fat based on only easily calculated body measurements (say, biceps and abdomen measurements);

Some Problems that Statistics can consider

- Find the relationship between the High School Math average grade and the average of the University Calc 1 and Calc 2 courses;

Note: Clearly, there is **no** exact formula relating the average of Calc1 and Calc2 grades to HS Math average grade. In statistics, we do not *find* exactly, but try to give a reasonable *Estimate*, even we assess how good is our Estimate;

- Calculate the percentage of body fat based on only easily calculated body measurements (say, biceps and abdomen measurements);
- Predict apartment prices based on various (given) characteristics of that apartment;

Some Problems that Statistics can consider

- Find the relationship between the High School Math average grade and the average of the University Calc 1 and Calc 2 courses;

Note: Clearly, there is **no** exact formula relating the average of Calc1 and Calc2 grades to HS Math average grade. In statistics, we do not *find* exactly, but try to give a reasonable *Estimate*, even we assess how good is our Estimate;

- Calculate the percentage of body fat based on only easily calculated body measurements (say, biceps and abdomen measurements);
- Predict apartment prices based on various (given) characteristics of that apartment;
- Check if the newly developed drug is more effective (in curing some disease) than the old one;

Some Problems that Statistics can consider

- Find the relationship between the High School Math average grade and the average of the University Calc 1 and Calc 2 courses;

Note: Clearly, there is **no** exact formula relating the average of Calc1 and Calc2 grades to HS Math average grade. In statistics, we do not *find* exactly, but try to give a reasonable *Estimate*, even we assess how good is our Estimate;

- Calculate the percentage of body fat based on only easily calculated body measurements (say, biceps and abdomen measurements);
- Predict apartment prices based on various (given) characteristics of that apartment;
- Check if the newly developed drug is more effective (in curing some disease) than the old one;
- Check if the drug is effective against Corona

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 YSU. 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 YSU. 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 YSU. 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 YSU. 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 YSU. 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 YSU. 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