

STATISTICS

ENG 3120

2023 - 2024 Spring Semester

 Assoc. Prof. Dr. Bora CANBULA

 www.canbula.com

 github.com/canbula/Statistics

 wn45g9v

Instructor

Assoc. Prof. Dr.
Bora CANBULA

Phone

0 (236) 201 21 08

Email

bora.canbula@cbu.edu.tr

Office Location

Dept. of CENG

Office C233

Office Hours

4 pm – 5 pm, Mondays

Course Overview

Statistics (Teams Code: wn45g9v)

We are going to learn both the mathematical foundations and real-world application of the statistics and the probability in this course. Focus of this course will be to provide the required background for a data science / machine learning course. Python is preferred as the programming language for the applications of this course.

Required Text

Probability And Statistics for Computer Scientists, CRC Press, *Michael Baron*

Introduction to Probability and Statistics, Elsevier, *Sheldon M. Ross*

Probability and Statistics for Engineers and Scientists, Brooks/Cole, *A.J. Hayter*

Course Materials

Python 3.x (Anaconda is preferred)

Jupyter Notebook from Anaconda

Pycharm from JetBrains / Visual Studio Code from Microsoft

Week	Subject	Week	Subject
1	Definitions of Descriptive Statistics	8	Linear Regression
2	Data, Sampling, and Variation	9	Linear Regression with Matrix Algebra
3	Visualization of Data	10	Regression with High Degree Polynomials
4	Measures of Central Tendency	11	Data Linearization and Transformation
5	Measures of Variation	12	Chi-Square and Goodness-of-Fit Tests
6	Measures for Multiple Variables	13	Central Limit Theorem
7	Box Plots and Outliers	14	Probability Distributions

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting the **data**.

Data is any kind of information.

Data are the actual values of the variable and can be categorical or numerical.

Population is the collection of people, things, or objects under study.

Sample is a subset of the population.

Statistic is a number that represents a property of the sample.

Parameter is a characteristic of the whole population that can be estimated by a statistic.

Variable is a characteristic or measurement that can be determined for each member of a population. They can be **dependent** or **independent**.

→ **Qualitative Variables** take on **Categorical** values.

→ **Quantitative Variables** take on **Numerical** values.

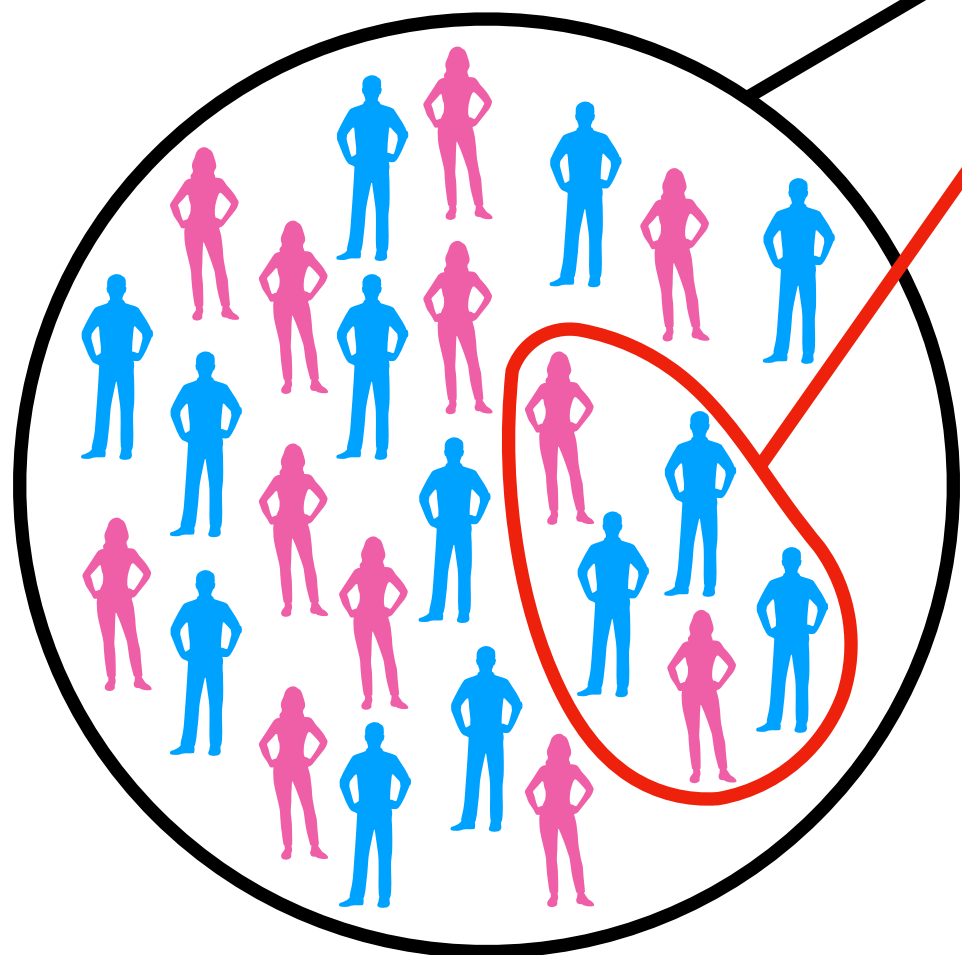
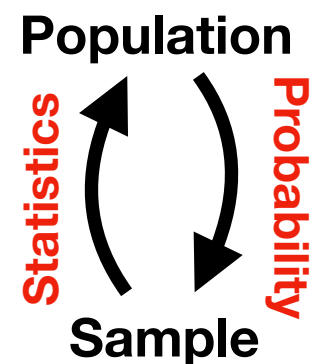
→ **Discrete Variables** take on finite number of values such as integers.

→ **Continuous Variables** take on infinite number of values such as real numbers.

Statistics

→ **Descriptive Statistics** is organizing and summarizing the data.

→ **Inferential Statistics** is drawing conclusions from good data.



good data == good sample

a good sample must be both random and representative

QUESTION








A study was conducted at our department to analyze the average GPA's of students who graduated last year. Match the key terms given below with the phrases that describes best.

A) Population **B)** Statistic **C)** Parameter **D)** Sample **E)** Variable **F)** Data

- ☒ **D** A group of students who graduated from our department last year
- ☒ **X** All students who attended last year
- ☒ **E** GPA of one student who graduated from our department last year
- ☒ **C** The average GPA of students who graduated from our department last year
- ☒ **A** All students who graduated from our department last year
- ☒ **F** 3.65, 2.80, 3.15, 3.90
- ☒ **B** _____

QUESTION

We plan on conducting a survey to our recent graduates to determine information on their yearly salaries. We randomly select 50 recent graduates and sent them questionnaires dealing with their present jobs. Of these 50, however, only 36 were returned. Suppose that the average of the yearly salaries reported was 415000 TL.

-  The population is: **Our all recent graduates**
-  The sample is: **36 recent graduates who returned to questionnaire**
-  The statistic is: **Yearly salary of 36 students**
-  The parameter is: **Yearly salary of our all recent graduates**
-  The variable is: **Yearly salary of one recent graduates**
-  Would we be correct in thinking that 415000 TL was a good approximation to the average salary level for all of our graduates? **No**
-  If your answer is no, can you think of any set of conditions relating to the group that returned questionnaires for which it would be a good approximation? **Suggest some questions**

QUESTION

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been in a malpractice lawsuit.

- ✍ The population is: **All medical doctors listed in the prof. directory**
- ✍ The sample is: **Selected 500 doctors**
- ✍ The statistic is: **The proportion of medical doctors in the sample**
- ✍ The parameter is: **The proportion of medical doctors in population**
- ✍ The variable is: **The number of medical doctors who have been**
- ✍ The data are: **Yes / No**

QUESTION

Determine the correct data type for the variables given below. Indicate whether quantitative data are continuous or discrete.

A) Numerical and discrete **B)** Numerical and continuous **C)** Categorical

- ☒ **A** The number of pairs of shoes you own
- ☒ **C** Gender
- ☒ **B** The distance from your home to university
- ☒ **A** The number of courses you take this semester
- ☒ **C** The brand of your mobile phone
- ☒ **B** Your weight
- ☒ **A** Number of correct answers on a quiz
- ☐ **?** Age