Data Analytics and Mining

Intro to Data Analytics, Mining, and Statistics

Data Analytics and Mining, 2024

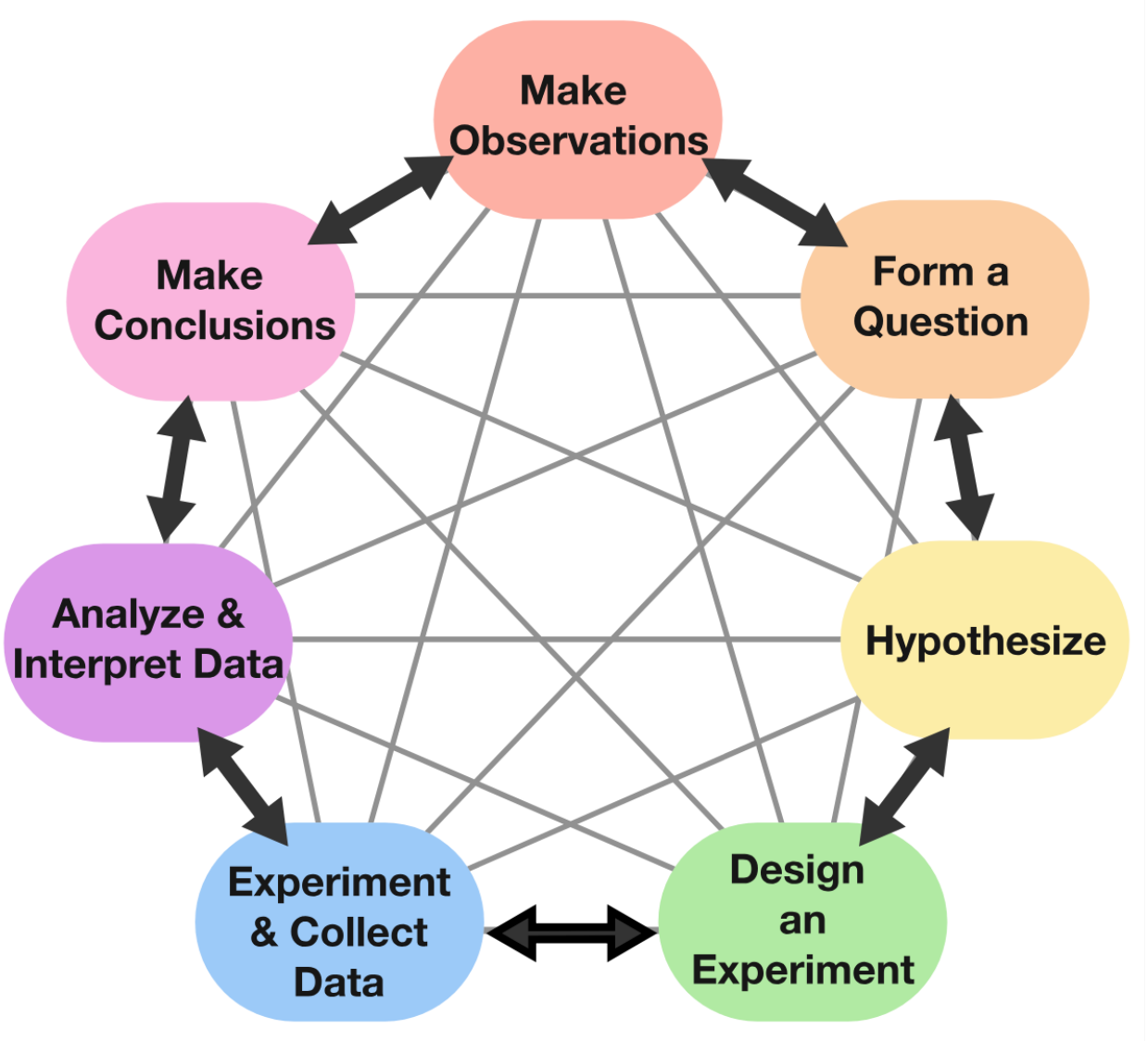
Majid Sohrabi
National Research University Higher School of Economics

October 04, 2024

Goals of the course

- To provide an understanding of the statistical concept and tests.
- To convey the principles of objectivity and logic that are essential for science and valuable for decision-making in everyday life.
- To provide a framework for a data investigation.
- To help acquire the skill of data manipulation via the job-market relevant software and tools (Python and its libraries)

Working with Data



What is statistics?

- The term statistics refers to a set of mathematical procedures for organizing, summarizing, and interpreting information.

- Statistics are used to organize and summarize the information so that the researcher can see what happened in the research study and can communicate the results to others.

- Statistics help the researcher to answer the questions that initiated the research by determining exactly what general conclusions are justified based on the specific results that were obtained.

Goals of the lectures?

- To get an idea of what data analytics, mining, and statistics actually refer to.
- To get familiar with the terminology.
- To get familiar with the theory of algorithms and tools we use in data analytics and mining. Such as machine learning methods using Python.

Population and samples

- A population is the set of all the individuals of interest in a particular study.

- A sample is a set of individuals selected from a population, usually intended to represent the population in a research study.

Population and samples

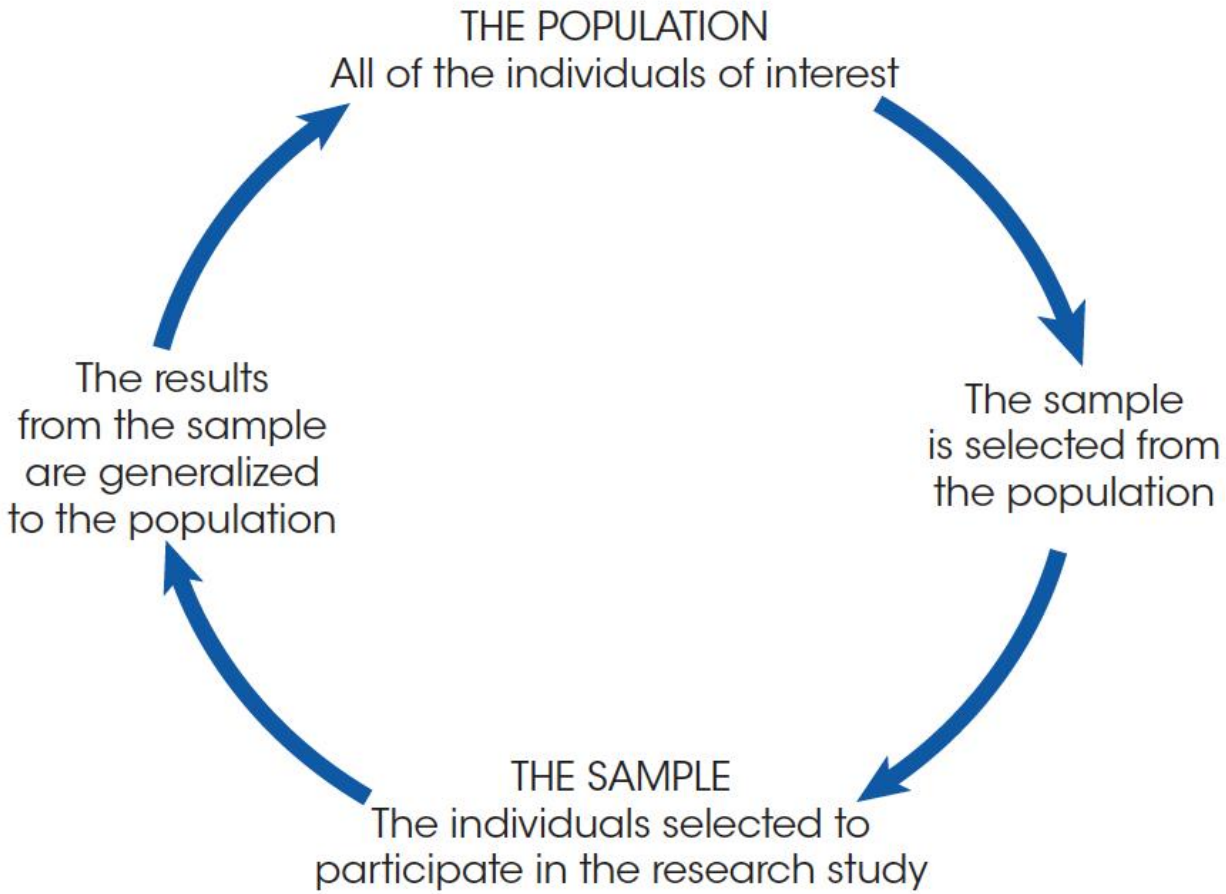


FIGURE 1.1
The relationship between a population and a sample.

Variables and data

- A variable is a characteristic or condition that changes or has different values for different individuals.

- **Data** (plural) are measurements or observations. A data set is a collection of measurements or observations. A **datum** (singular) is a single measurement or observation and is commonly called a score or raw score.

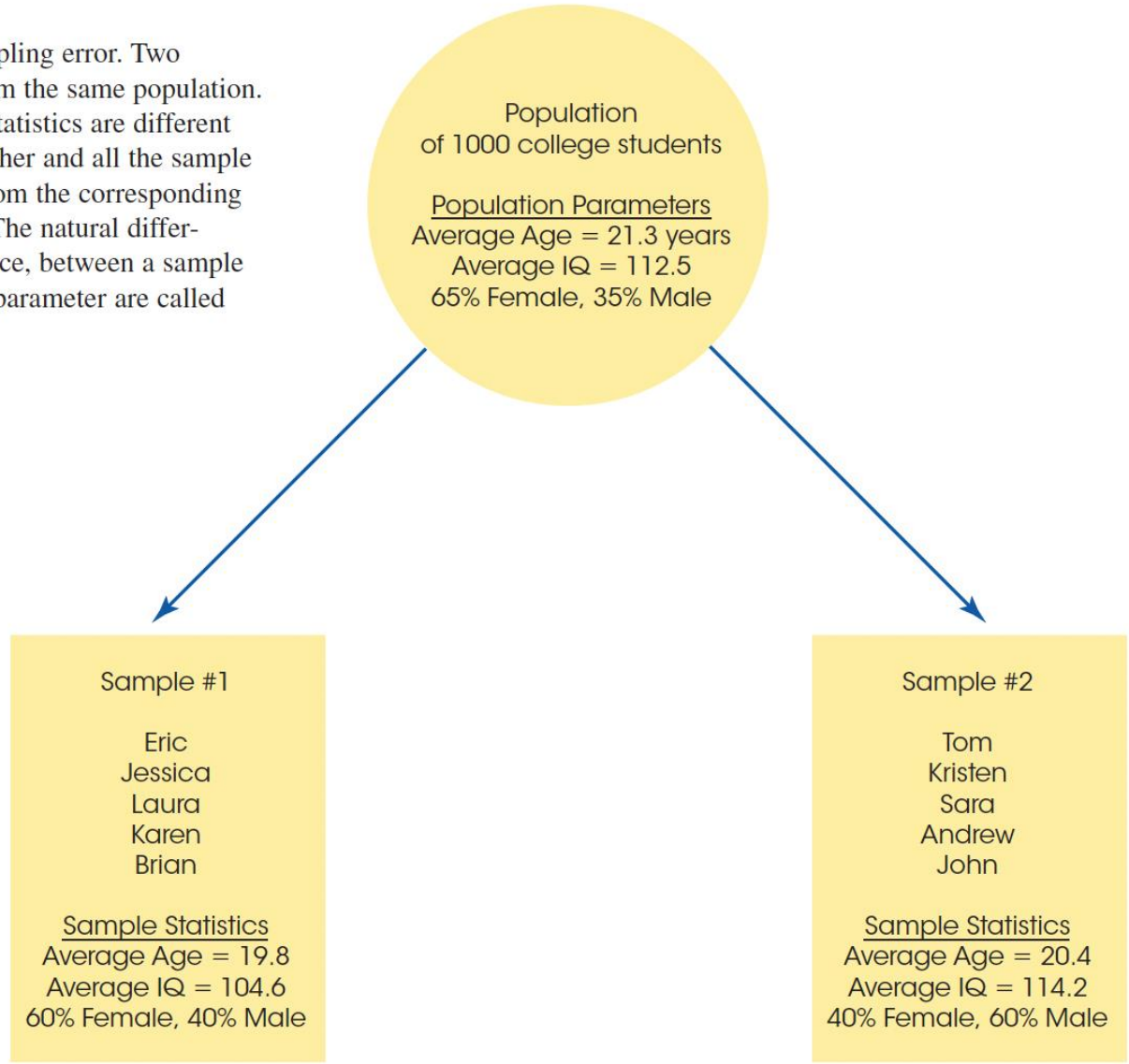
Parameters and statistics

- A **parameter** is a value, usually a numerical value, that describes a population. A parameter is usually derived from measurements of the individuals in the population.

- A **statistic** is a value, usually a numerical value, that describes a sample. A statistic is usually derived from measurements of the individuals in the sample.

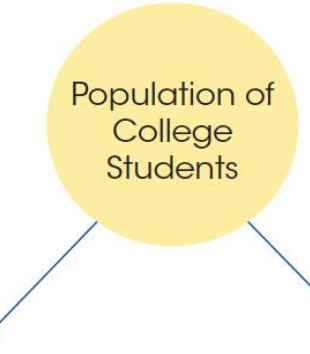
Parameters and statistics

FIGURE 1.2
A demonstration of sampling error. Two samples are selected from the same population. Notice that the sample statistics are different from one sample to another and all the sample statistics are different from the corresponding population parameters. The natural differences that exist, by chance, between a sample statistic and population parameter are called sampling error.



Statistics in research context

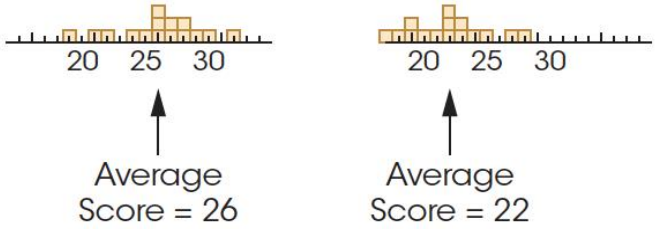
Step 1
Experiment:
Compare two studying methods



Data
Test scores for the students in each sample

Sample A Read from printed pages			Sample B Read from computer screen		
25	26	28	20	22	27
27	21	27	23	17	23
30	28	24	25	28	21
19	23	26	22	19	22
29	26	22	18	24	19

Step 2
Descriptive statistics:
Organize and simplify



Step 3
Inferential statistics:
Interpret results

The sample data show a 4-point difference between the two methods of studying. However, there are two ways to interpret the results.

1. There actually is no difference between the two studying methods, and the sample difference is due to chance (sampling error).
2. There really is a difference between the two methods, and the sample data accurately reflect this difference.

The goal of inferential statistics is to help researchers decide between the two interpretations.

experimental

Learning check

1. A researcher is interested in the sleeping habits of American college students. A group of 50 students is interviewed and the researcher finds that these students sleep an average of 6.7 hours per day. For this study, the average of 6.7 hours is an example of a(n) _____.

a. parameter

b. statistic

c. population

d. sample

Learning check

2. A researcher is curious about the average IQ of registered voters in the state of Florida. The entire group of registered voters in the state is an example of a _____.

a. sample

b. statistic

c. population

d. parameter

Research Methods

- Individual variables: descriptive research.

- Relationships between variables
  - Correlational method
  - Comparing Two (or More) groups of scores: experimental and nonexperimental methods

Correlational method

(a)

Student	Wake-up Time	Academic Performance
A	11	2.4
B	9	3.6
C	9	3.2
D	12	2.2
E	7	3.8
F	10	2.2
G	10	3.0
H	8	3.0

(b)

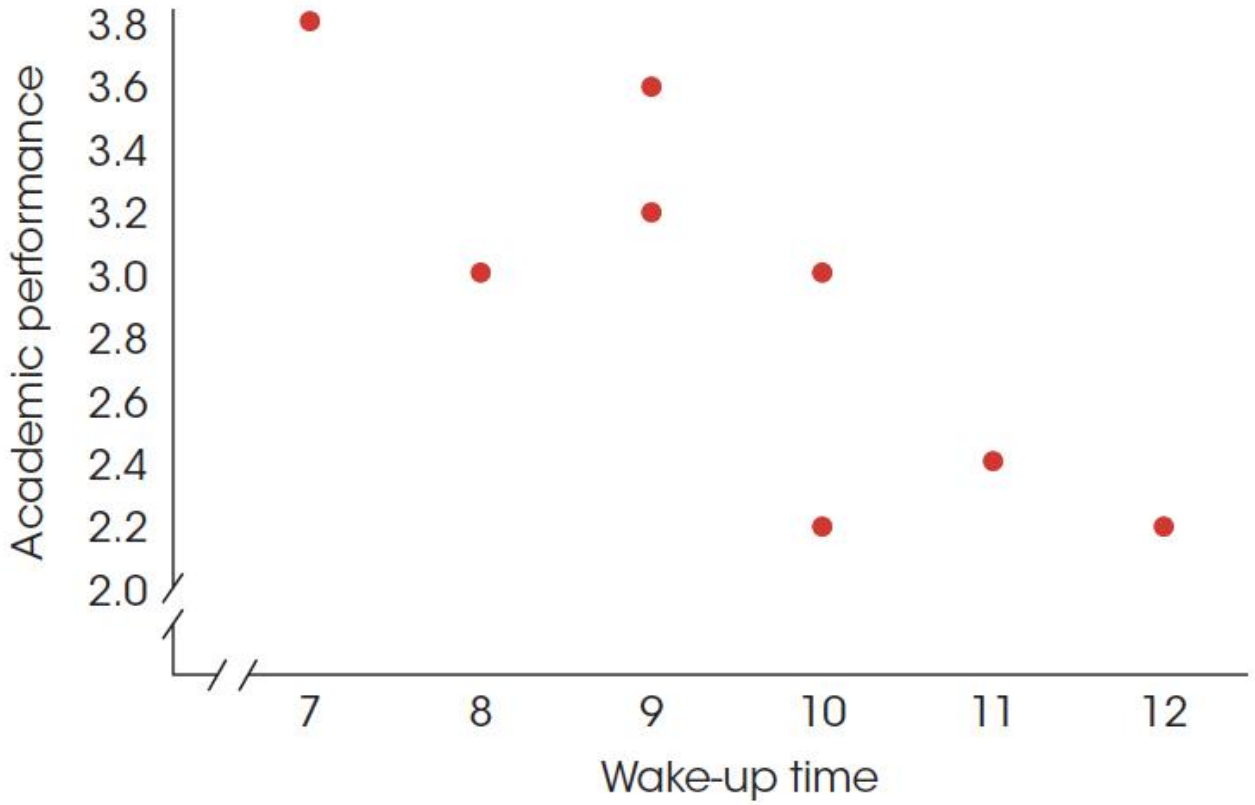


FIGURE 1.4

One of two data structures for evaluating the relationship between variables. Note that there are two separate measurements for each individual (wake-up time and academic performance). The same scores are shown in a table (a) and in a graph (b).

Experimental and nonexperimental methods

- In the **experimental method**, one variable is manipulated while another variable is observed and measured. To establish a cause-and-effect relationship between the two variables, an experiment attempts to control all other variables to prevent them from influencing the results.

- In the **nonexperimental** (observational) method we use data collected in other ways to determine the relationship between variables.

Experimental and nonexperimental methods

- The **independent variable** is the variable that is manipulated by the researcher in the experiment or the one observed that might influence our variable of interest.

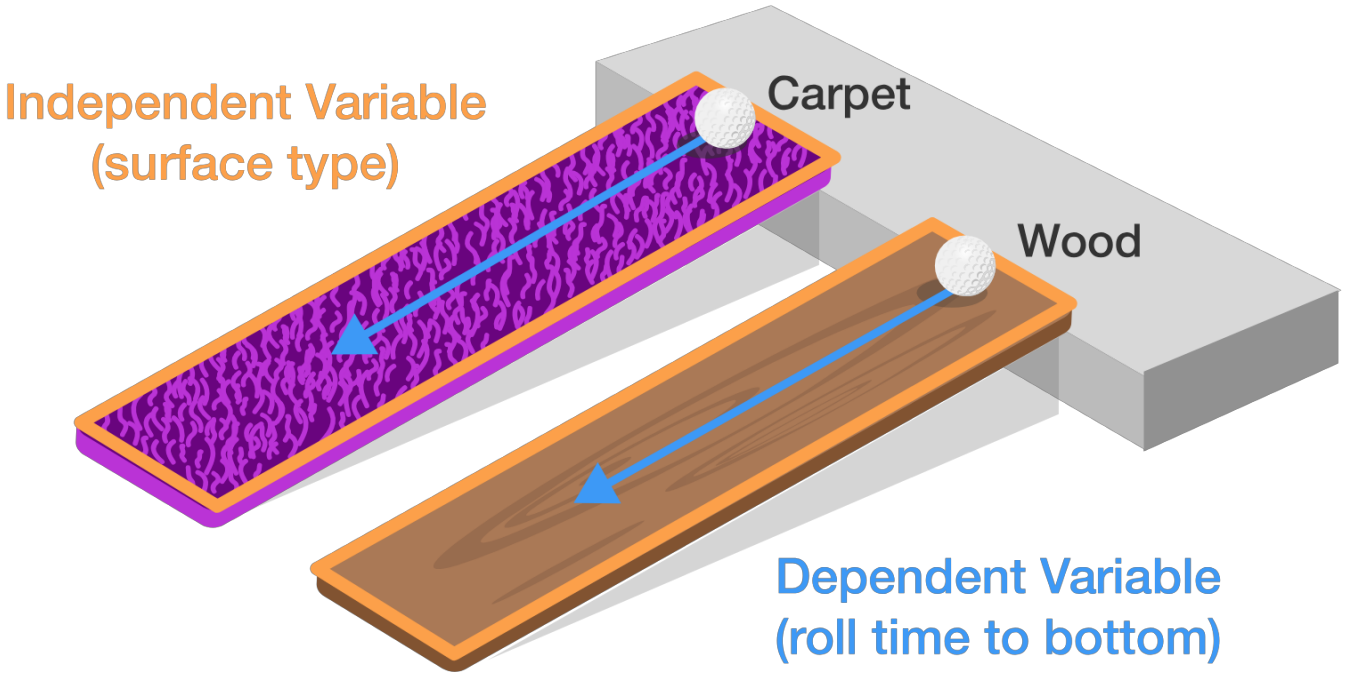
- The **dependent variable** is the one that is observed to assess the effect of the treatment or the one we hypothetize to be affected by others.

- The **control variables** are the variables that researchers seek to keep constant when conducting research to isolate the influence of the dependent variable.

Experimental method. Example

Research question: Does the type of the surface affect the speed of the ball?

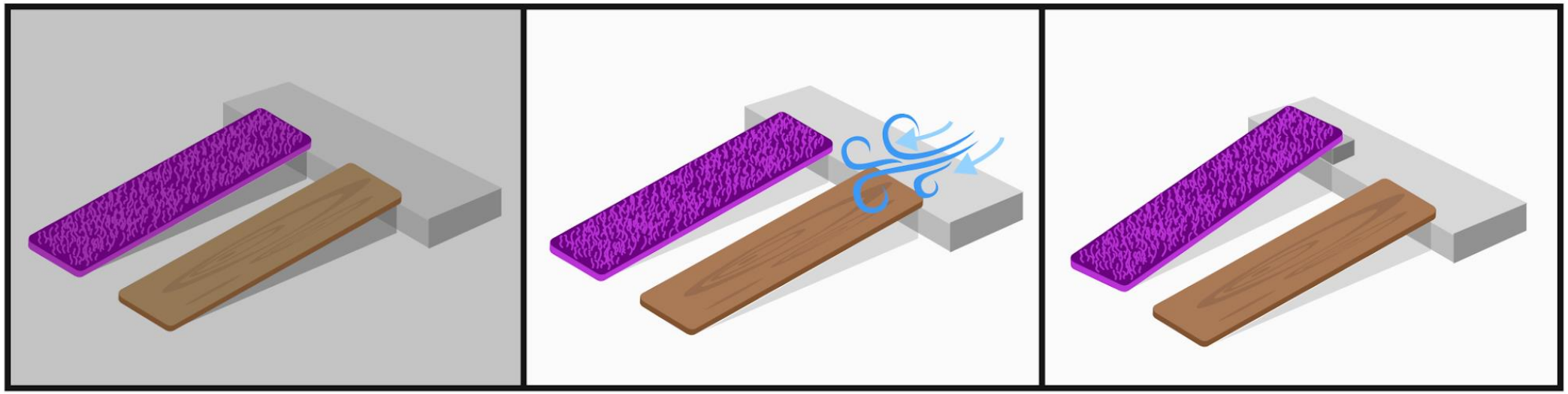
Hypothesis: All things are equal, the speed of ball will be higher on the surface with less resistance.



Experimental method. Example

What else can affect the speed?

Amount of light in the room?
Wind?
Angle?



Experimental method. Example

Dependent variable:

Independent variable:

Control variables:

Experimental method. Example

Dependent variable: speed of ball

Independent variable: surface type

Control variables: ball type, humidity, angle, wind speed, etc.

Experimental method. Example

Ball type	Surface type	Ball speed	Wind speed	Humidity
table tennis ball	wood	3 m/s	0	24%
table tennis ball	carpet	2 m/s	0	24%

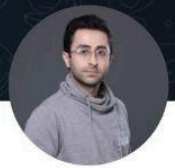
Summary

- What are statistics and the basic terminology
- What are the population and samples
- What is the difference between inferential and descriptive statistics
- What major research methods are
- Looked at the experiment example

Thank you!

Majid Sohrabi

msohrabi@hse.ru



@MSOHRABI_CS