STATISTICS

ENG 3120

2023 - 2024 Spring Semester

Assoc. Prof. Dr. Bora CANBULA

www.canbula.com

github.com/canbula/Statistics

wn45g9v

Manisa Celal Bayar University *Department of Computer Engineering*

Syllabus

STATISTICS

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

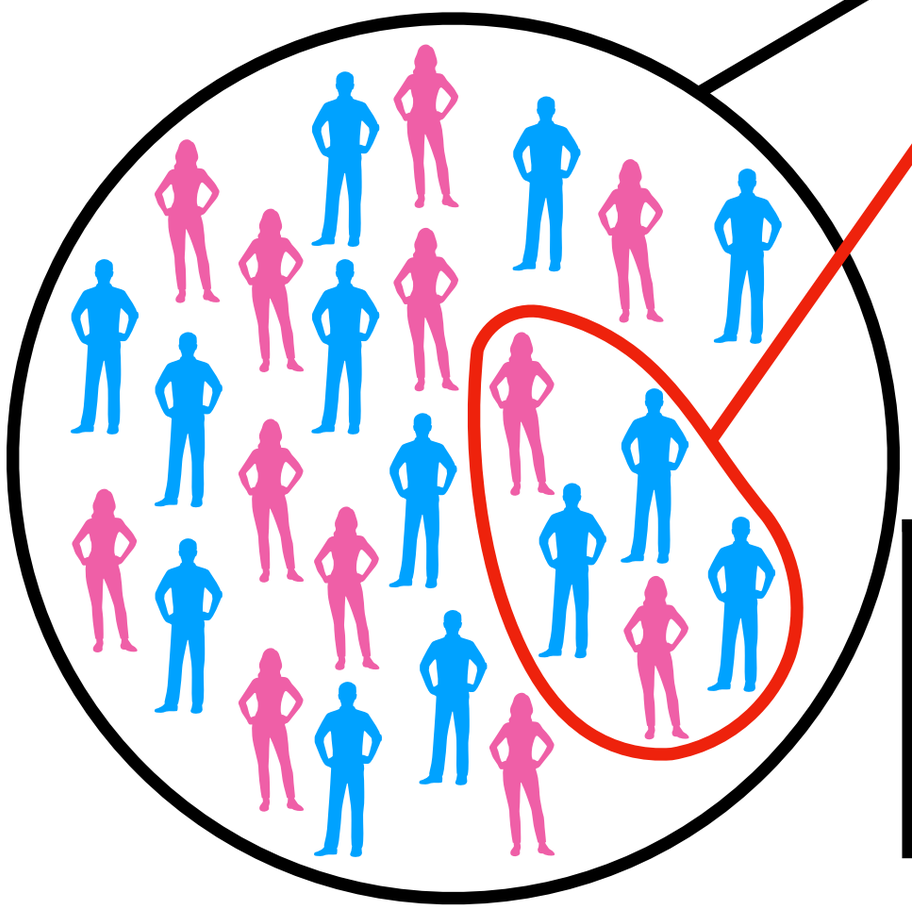
Definitions of Descriptive Statistics

STATISTICS

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.



good data == good sample

a good sample must be both random and representative

Statistics

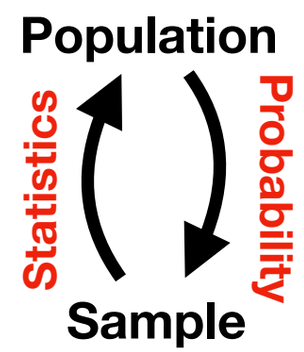
- **Descriptive Statistics** is organizing and summarizing the data.
- **Inferential Statistics** is drawing conclusions from good data.

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.

Definitions of Descriptive Statistics

STATISTICS

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 _____

Definitions of Descriptive Statistics

STATISTICS

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**

Definitions of Descriptive Statistics

STATISTICS

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**

Definitions of Descriptive Statistics

STATISTICS

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

Sampling

Population (N) ⟶ **Sample (n)**

BEST

HARDEST

Simple Random Sampling

Every member of the population has an equal chance to be in the sample.

Stratified Sampling

The population is split into non-overlapping groups, which is called strata, then simple random sampling is applied to each group.

Cluster Sampling

The population is divided into groups (clusters), then some of the clusters are randomly selected.

Systematic Sampling

The sample is constructed with every n^{th} individual from the population.

Convenience Sampling

The sample is constructed with easily obtained members of the population.

WORST

EASIEST

Determine the type of sampling used in the following examples:

- A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.

  Stratified Sampling
- A pollster interviews all human resource personnel in five different high tech companies.

  Cluster Sampling
- A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.

  Stratified Sampling
- A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

  Systematic Sampling
- A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.

  Simple Random Sampling
- A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

  Convenience Sampling

Sampling

Population (N) ⟶ **Sample (n)**

BEST

HARDEST

Simple Random Sampling

Every member of the population has an equal chance to be in the sample.

```
cities = [
    "Adana",
    "Adıyaman",
    "Afyonkarahisar",
    "Ağrı",
    # ...
    "Kilis",
    "Osmaniye",
    "Düzce",
]
```

```
import sys

sys.path.append(".")


from Week02 import data

print(data.cities)
```

```
import random
```

```
[
    "betavariate",
    "binomialvariate",
    "choice",
    "choices",
    "expovariate",
    "gammavariate",
    "gauss",
    "getrandbits",
    "getstate",
    "lognormvariate",
    "normalvariate",
    "paretovariate",
    "randbytes",
    "randint",
    "random",
    "randrange",
    "sample",
    "seed",
    "setstate",
    "shuffle",
    "triangular",
    "uniform",
    "vonmisesvariate",
    "weibullvariate",
]
```

```
Help on method sample i

sample(population, k, *
    Chooses k unique ra

    Returns a new list
    leaving the origina
    in selection order
    samples.  This allo
    into grand prize an

Members of the popu
population contains
selection in the sa

Repeated elements c
counts parameter.

    sample(['red',

is equivalent to:

    sample(['red',

To choose a sample
population argument
for sampling from a

    sample(range(10
```

```
import sys

sys.path.append(".")


from Week02 import data
import random


def simple_random_sampling(data, n):
    return random.sample(data, n)


sample = simple_random_sampling(data.cities, 10)
print(sample)
```

WORST

EASIEST

Sampling

Population (N) ⟶ **Sample (n)**

BEST

Stratified Sampling

The population is split into non-overlapping groups, which is called strata, then simple random sampling is applied to each group.

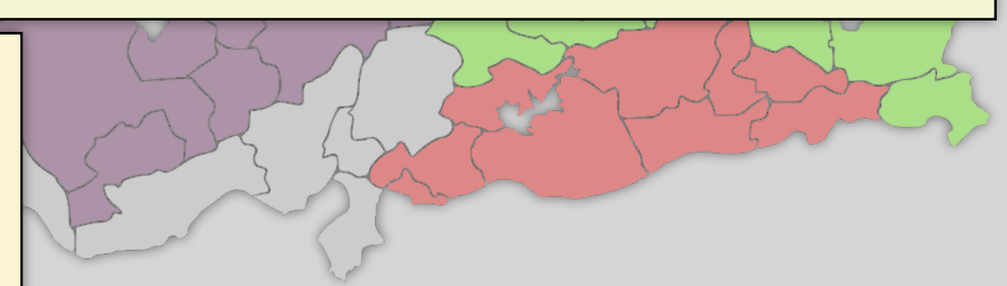
```python
regions = [
    "Marmara",
    "İç Anadolu",
    "Ege",
    "Akdeniz",
    "Karadeniz",
    "Doğu Anadolu",
    "Güneydoğu Anadolu",
]
```

```python
def stratified_sampling(data, n, strata):
    sample = []
    for key in strata:
        sample += random.sample(data[key], n)
    return sample


sample = stratified_sampling(data.cities_by_region, 3, data.regions)
print(sample)
```

Fix this!

```python
cities_by_region = {
    "Marmara": ["Edirne", "Kırklareli", '
    "İç Anadolu": ["Aksaray", "Ankara", '
    "Ege": ["İzmir", "Manisa", "Aydın", '
    "Akdeniz": ["Adana", "Osmaniye", "Ant
    "Karadeniz": ["Rize", "Trabzon", "Art
    "Doğu Anadolu": ["Ağrı", "Ardahan", '
    "Güneydoğu Anadolu": ["Adıyaman", "Ba
}
```



HARDEST

WORST

EASIEST

Sampling

Population (N) ⟶ Sample (n)

BEST

HARDEST

Cluster Sampling

The population is divided into groups (clusters), then some of the clusters are randomly selected.

```python
regions = [
    "Marmara",
    "İç Anadolu",
    "Ege",
    "Akdeniz",
    "Karadeniz",
    "Doğu Anadolu",
    "Güneydoğu Anadolu",
]
```

```python
def stratified_sampling(data, n, strata):
    sample = []
    for key in strata:
        sample += random.sample(data[key], n)
    return sample


sample = stratified_sampling(data.cities_by_region, 3, data.regions)
print(sample)
```

Fix this!

```python
cities_by_region = {
    "Marmara": ["Edirne", "Kırklareli", '
    "İç Anadolu": ["Aksaray", "Ankara", '
    "Ege": ["İzmir", "Manisa", "Aydın", '
    "Akdeniz": ["Adana", "Osmaniye", "Ant
    "Karadeniz": ["Rize", "Trabzon", "Art
    "Doğu Anadolu": ["Ağrı", "Ardahan", '
    "Güneydoğu Anadolu": ["Adıyaman", "Ba
}
```

```python
def cluster_sampling(data, n, clusters):
    picked_clusters = random.sample(clusters, n)
    sample = []
    for cluster in picked_clusters:
        sample += data[cluster]
    return sample
```

Duplicates?

WORST

EASIEST

Sampling

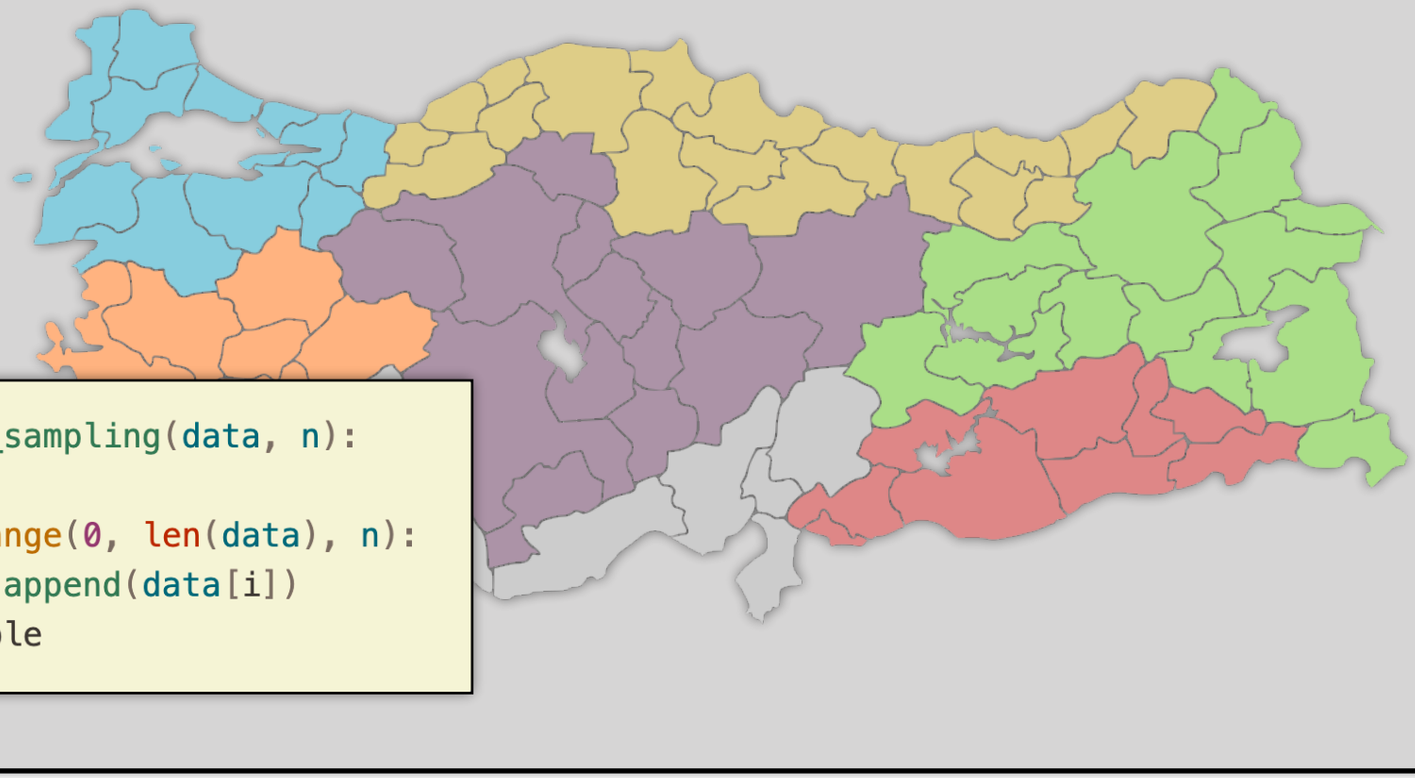
Population (N) ⟶ **Sample (n)**

BEST

HARDEST

Systematic Sampling

The sample is constructed with every n^{th} individual from the population.



```python
def systematic_sampling(data, n):
    sample = []
    for i in range(0, len(data), n):
        sample.append(data[i])
    return sample
```

WORST

EASIEST

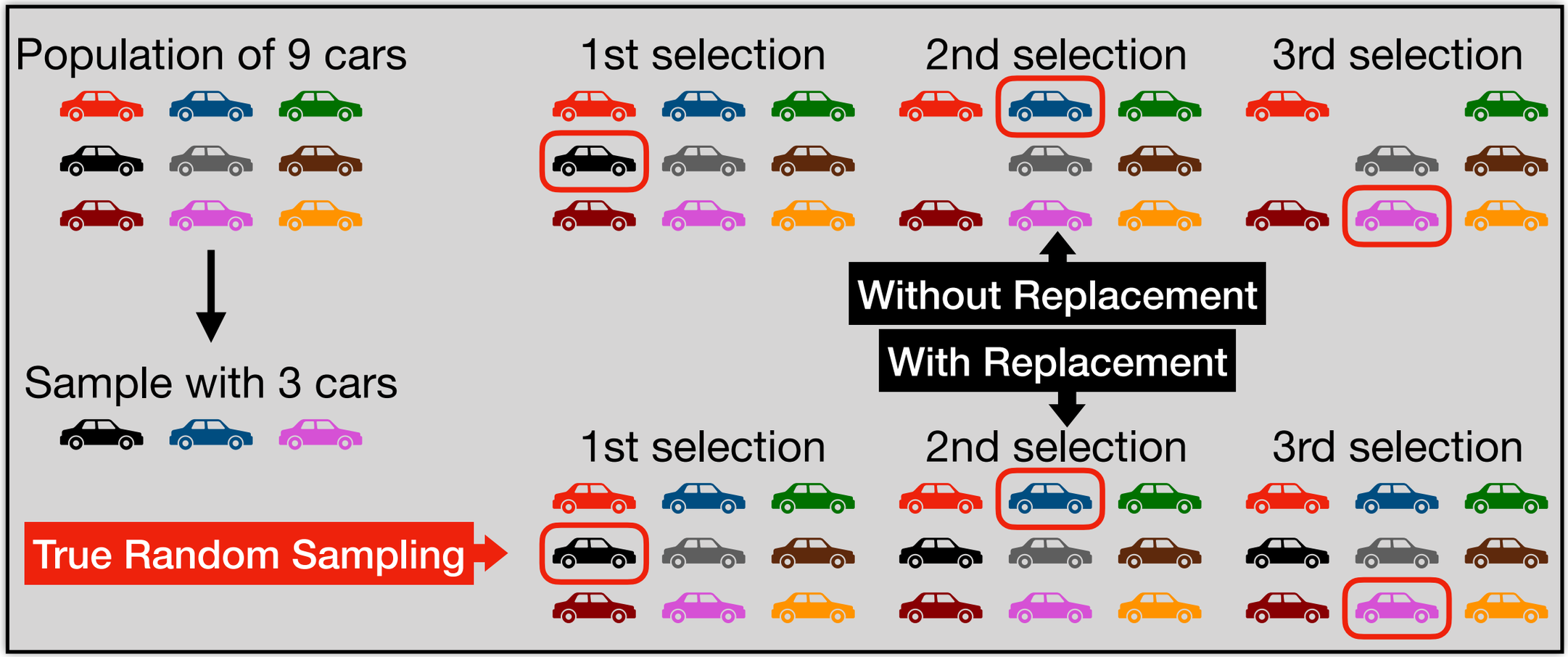
Sampling

Population (N) ⟶ **Sample (n)**

BEST

Simple Random Sampling

Every member of the population has an equal chance to be in the sample.



WORST

HARDEST

EASIEST

IMPLEMENT THIS FEATURE AS A HOMEWORK

Weighted Simple Random Sampling w/ Replacement Support

Name of Your File

Week03/weighted_firstname_lastname.py

Name of the Function

weighted_srs

Input Parameters

- data [list]: population
- n [int]: sample size
- weights [list]: weights for members of population
- with_replacement [bool]: flag for true random sampling

Other Rules

- Using maximum 10 lines of codes is allowed
- Using any modules, other than random, is forbidden

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. The correct statistical procedures that can be used with a data set is specified with this level.

	Difference	Order	Similar Intervals	Natural Zero
Less Information				
Categorical Data — **Nominal**: Nominal level represents the categories that **cannot** be put in any order	✓	✗	✗	✗
Categorical Data — **Ordinal**: Ordinal level represents the categories that **can** be put in a order	✓	✓	✗	✗
Numerical Data — **Interval**: Interval level has a definite ordering, and **distances between values are equal and meaningful**	✓	✓	✓	✗
Numerical Data — **Ratio**: Ratio level provides the most information: order, fixed scale, and also a **natural zero**	✓	✓	✓	✓
More Information				

Levels of Measurement

STATISTICS

Determine the type of measure scale used in the following examples:

- Letter Grades: AA, BA, BB, CB, CC, …

  Ordinal Scale

- The number of students in a classroom

  Ratio Scale

- The dates 1997, 2004, 2020, …

  Interval Scale

- Political outlook: extreme left, left-of-center, right-of-center, extreme right

  Nominal Scale

- Turkish Republic identification number

  Nominal Scale

Measures of Central Tendency

STATISTICS

Measures of central tendency are used to determine the center of a distribution of data. It is used to find a single score that is the most representative of an entire data set.

Mean is simply the arithmetic **average** of the data observations.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Median is the value in the **middle** of the ordered data points.

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & n : \text{odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & n : \text{even} \end{cases}$$

Mode is the value with the **highest frequency** of the data set.

QUESTION

If a constant c is added to each x_i in a sample, yielding $y_i = x_i + c$ how do the sample mean and median of the y_is relate to the mean and median of the x_is?

If each x_i is multiplied by a constant c, yielding $y_i = cx_i$, answer the question again.

HW

Week05/shifted_firstname_lastname.py

Function **`shifted`** calculates the difference between the mean and median in percentage.