Oct 12, 2023

Engineering Statistics

Week 1: Fundamental Elements of Statistics

Evaluation

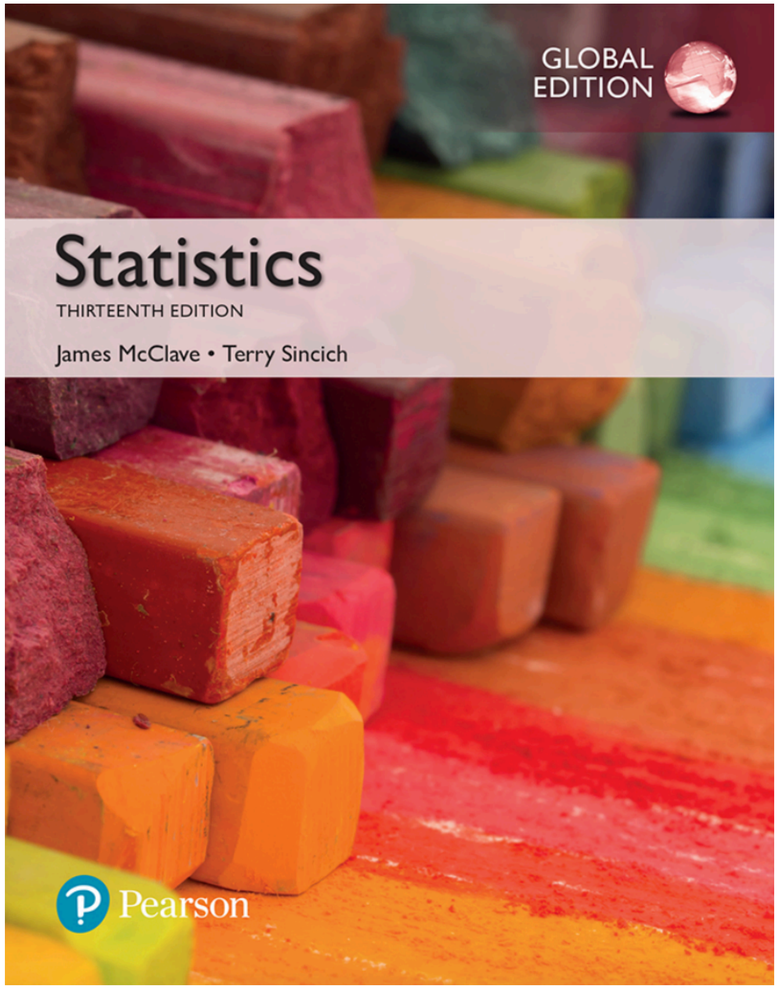
The evaluation of this course is done

Midterm exam on Nov 15, 2023 at 18:00

Final exam on Jan 10, 2024 at 18:00

Reference

McClave and Sincich (2018) “Statistics”, Thirteenth Edition, Pearson.



What is Statistics?

Statistics is the science that deals:

- collection,
- classification,
- analysis,
- interpretation of the data.

Data

- **Data** are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

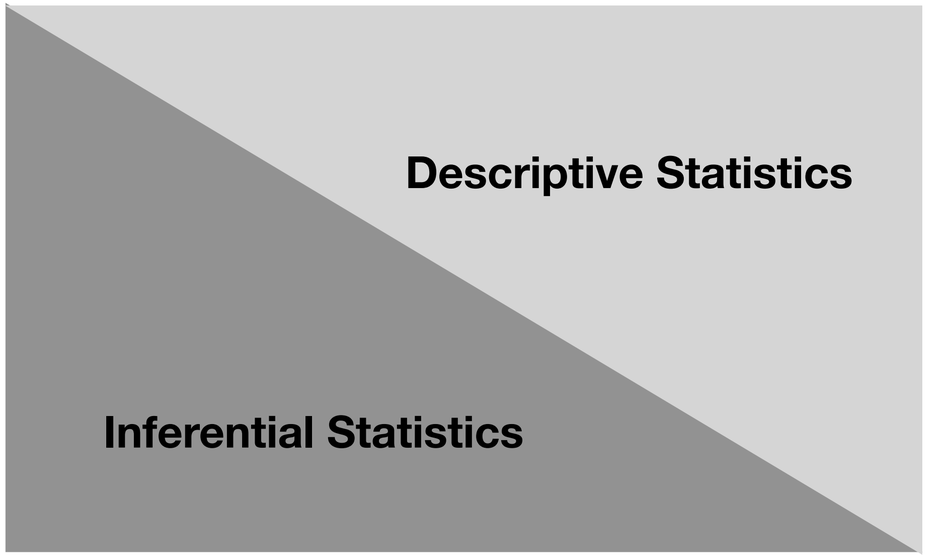
- All the data collected in a particular study are referred to as the **data set** for the study.

- Measurements collected on each variable for every element in a study provide the data. The set of measurements obtained for a particular observational unit is called as an **observation**.

Type of Statistical Applications

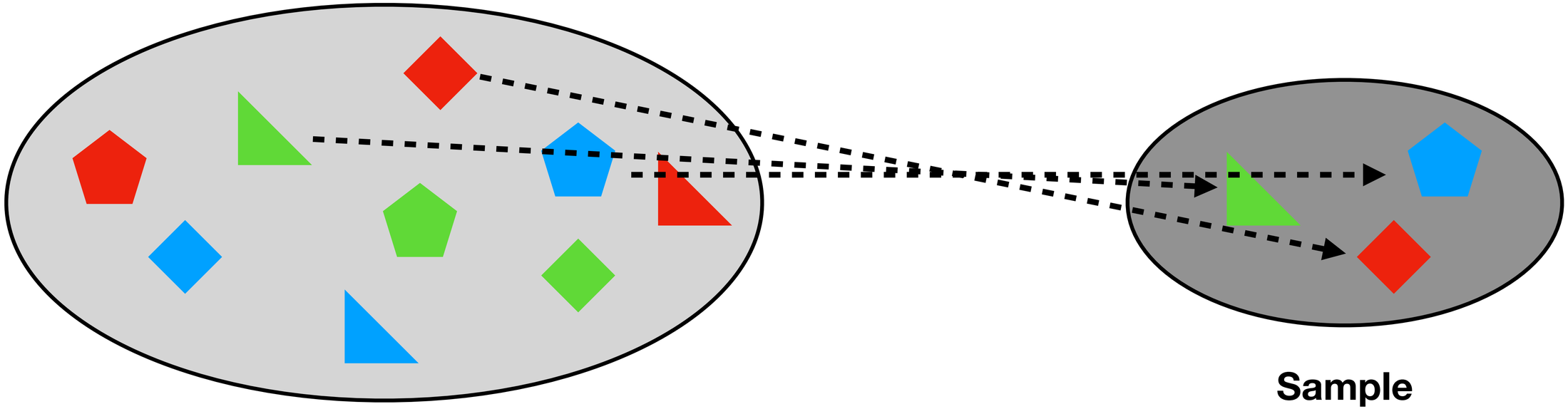
utilizes **the sample data:**
- to make estimates,
- decisions,
- predictions

or other generalizations about a larger set of data.



utilizes **numerical and graphical methods:**
- to look for patterns,
- to summarize the information revealed,
- to present the information in the dataset

Fundamental Elements of Statistics



Sample

Population

An observational (experimental) unit is an object about which we collect data.

A population is a set of all units that we are interested in studying.

A sample is a subset of the units of a population.

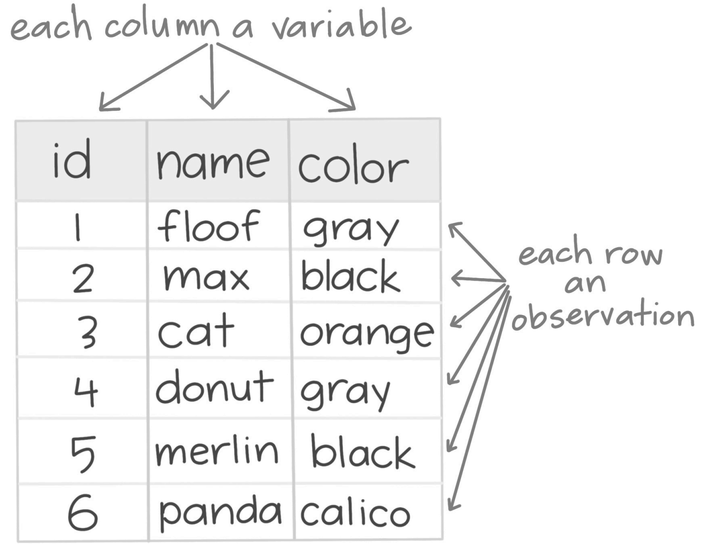
Fundamental Elements of Statistics

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



A **variable** is a characteristics or property of an individual observational unit in the population.

Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

Application

Assume that we are working on the stability of perovskite materials and 10 materials are randomly selected. Data contains **atomic volume, material composition, number of elements, and etc.**

Material composition	Atomic volume (cm^3/mol)	Number of elements	Energy above hull (meV/ atom)
Mg8Fe8O24	10.53500	3	636.33932
La1Y4Zn3Ni8O24	11.26562	5	291.95556
La1Y4Zn3Co8O24	11.32062	5	168.46496
La5Zn3Ni8O24	11.49812	4	304.21610
La1Pr4Zn3Ni8O24	11.51562	5	304.21610
La1Y4Zn3Fe8O24	11.52063	5	158.36741
La5Zn3Co8O24	11.55312	4	181.65975
La1Pr4Zn3Co8O24	11.57062	5	172.42020
La1Y4Mn8Zn3O24	11.66563	5	121.20444
La5Zn3Fe8O24	11.75313	4	176.19335

1. **Identify the observational unit for this study**
2. **Identify the population and sample**
3. **Identify the variables**

Paper of data: https://doi.org/10.1016/j.commatsci.2018.04.033

Statistical Inference

A statistical inference is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

We use the information contained in the smaller set of population to learn about the population. In short, **we observe sample but are interested in population**.

Application

Assume that we are working on the stability of perovskite materials. 10 perovskite materials are randomly selected. Data contains **atomic volume, material composition,** and **number of elements.**

Material composition	Atomic volume (cm^3/mol)	Number of elements	Energy above hull (meV/ atom)
Mg8Fe8O24	10.53500	3	636.33932
La1Y4Zn3Ni8O24	11.26562	5	291.95556
La1Y4Zn3Co8O24	11.32062	5	168.46496
La5Zn3Ni8O24	11.49812	4	304.21610
La1Pr4Zn3Ni8O24	11.51562	5	304.21610
La1Y4Zn3Fe8O24	11.52063	5	158.36741
La5Zn3Co8O24	11.55312	4	181.65975
La1Pr4Zn3Co8O24	11.57062	5	172.42020
La1Y4Mn8Zn3O24	11.66563	5	121.20444
La5Zn3Fe8O24	11.75313	4	176.19335

1. **Describe the population**
2. **Describe the variable of interest**
3. **Describe the sample**
4. **Describe the inference**

Data Types

The type of the data can be classified in two ways:

- Qualitative or Quantitative
- Scale of measurement

Data Types

Quantitative

- countable or measurable, relating to numbers
- tell us how many, how much, or how often

Qualitative

- descriptive, relating to words and language
- describes certain attributes, and help us to understand the “why” or “how” behind certain behaviors.

Scales of Measurements

1. Nominal scale
2. Ordinal scale
3. Interval scale
4. Ratio scale

Nominal scale

When the data for a variable **consist of labels or names** used to identify an attribute of the element, the scale of measurement is considered a nominal scale.

Example: Material composition, material type

Ordinal scale

The scale of measurement for a variable is called an ordinal scale if the data **exhibit the properties of nominal data and the order or rank** of the data is meaningful.

Example: Stability of a material (low, moderate, high)

Interval scale

- The scale of measurement for a variable is an interval scale if the data **have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure**. Interval data are always numeric.

  Example: Stability of three materials (A, B, and C) are measured as 500, 600, and 700 meV/Atom, respectively. We can say that material B is 600 - 500 = 100 meV/Atom more stabil than material A, or the material A is (500 - 600 = -100) 100 meV/Atom less stabil than material B.

- Zero is a point on interval scale, but it does not mean the absence of the condition.

  Example: Temperature takes the value of 0 and it means a level of temperature.

Ratio scale

- The scale of measurement for a variable is a ratio scale if the data have all the properties of interval data and the ratio of two values is meaningful. **This scale requires that a zero value be included to indicate that nothing exists for the variable at the zero point.**

  Example: Assume that stability of the materials A and B are measured as 500, and 1000 meV/Atom, respectively. We can say that material B is 1000/500 = 2 times more stabil than material A.

- **The difference between interval and ratio scale is that ratio scale never fall below zero.**

Application

Assume that we are working on the stability of perovskite materials. Data contains **atomic volume, material composition,** and **number of elements.**

Material composition	Atomic volume (cm^3/mol)	Number of elements	Energy above hull (meV/atom)
Mg8Fe8O24	10.53500	3	636.33932
La1Y4Zn3Ni8O24	11.26562	5	291.95556
La1Y4Zn3Co8O24	11.32062	5	168.46496
La5Zn3Ni8O24	11.49812	4	304.21610
La1Pr4Zn3Ni8O24	11.51562	5	304.21610
La1Y4Zn3Fe8O24	11.52063	5	158.36741
La5Zn3Co8O24	11.55312	4	181.65975
La1Pr4Zn3Co8O24	11.57062	5	172.42020
La1Y4Mn8Zn3O24	11.66563	5	121.20444
La5Zn3Fe8O24	11.75313	4	176.19335

- **Define the variables measured and classify them as quantitative or qualitative.**
- **Classify the categorical variables as ordinal or nominal.**

Data Collection Methods

Designed Experiment 🎨

- the researcher control over the characteristics of the observational units sampled.
- these experiments typically involve a group observational units that are assigned the treatment and control (untreated) group.

Observational Study 🔍

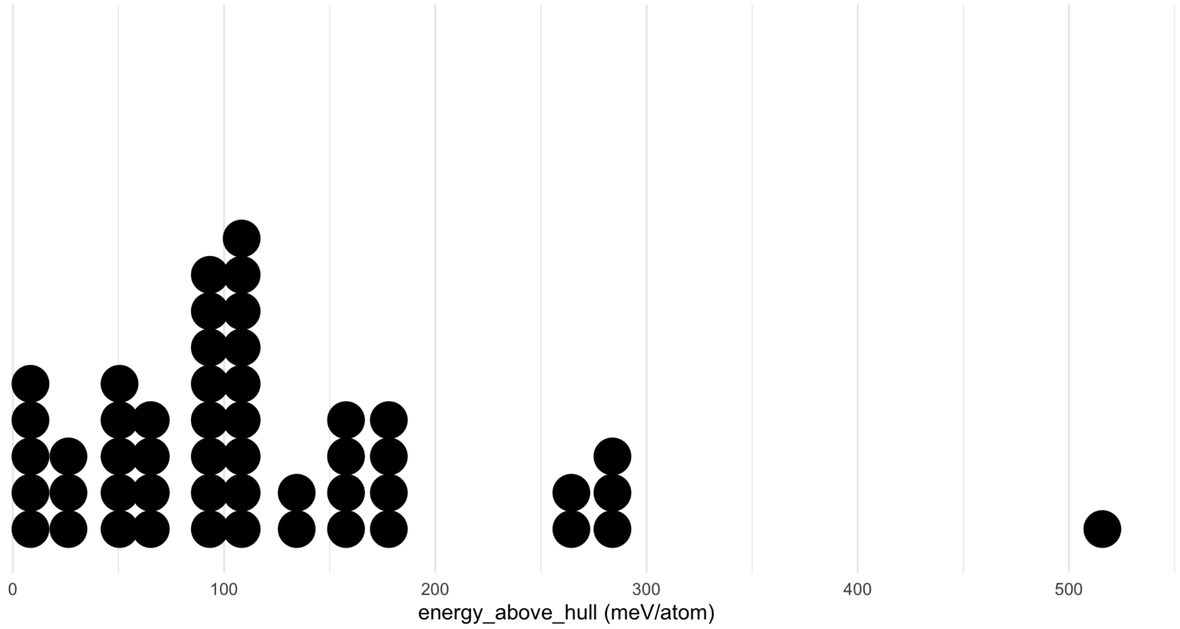
- the observational units sampled are observed in their natural setting.
- No attempt is made to control the characteristics of the observational units sampled.

Describing Quantitative Data

- dot plot
- stem-and-leaf plot
- histogram
- …

Dot plots

- The numerical value of each quantitative measurement in the data set is represented by a dot on a horizontal scale.
- When data values repeat, the dots are placed above one another vertically.



Stem-and-leaf plots

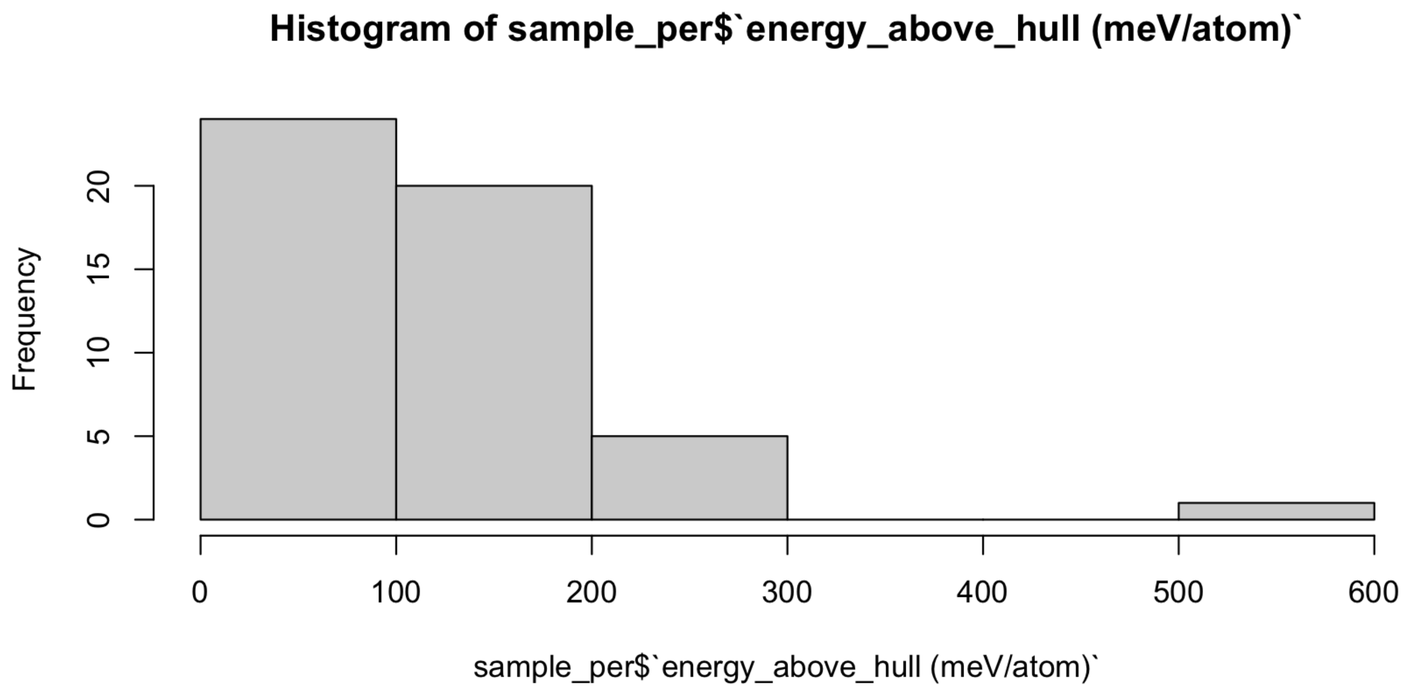
It gives a quick picture of the shape of a distribution while including the actual numerical values in the plot.

It works best for small number of observations that are all greater than zero.

```
0 | 011122334455666778999999
1 | 000111111113456667888
2 | 67889
3 |
4 |
5 | 2
```

Histogram

Histograms are used to display either the frequency or relative frequency of the measurements falling into the class intervals.



Pros and Cons of the Plots

- Histograms provide good visual descriptions of data sets - particularly very large ones - they do not let us identify individual measurements.
- In contrast, each of the original measurements is visible to some extent in a dot plot and is clearly visible in stem-and-leaf plot.
- The stem-and-leaf plot arranges the data in ascending order, so it is easy to locate the individual measurements.

Describing Qualitative Data

- bar plot
- pie chart
- …

Frequency Distribution

A frequency distribution is a tabular summary of data showing **the number (frequency) of items** in each several non-overlapping classes.

Key terms:

Class: one of the categories into which qualitative data can be classified.

Class Frequency: the number of observations in the data set that fall into a particular class.

Application

Assume that we are working on the stability of perovskite materials. Data contains **atomic volume, material composition,** and **number of elements.**

Material composition	Atomic volume (cm^3/mol)	Number of elements	Energy above hull (meV/atom)
Mg8Fe8O24	10.53500	3	636.33932
La1Y4Zn3Ni8O24	11.26562	5	291.95556
La1Y4Zn3Co8O24	11.32062	5	168.46496
La5Zn3Ni8O24	11.49812	4	304.21610
La1Pr4Zn3Ni8O24	11.51562	5	304.21610
La1Y4Zn3Fe8O24	11.52063	5	158.36741
La5Zn3Co8O24	11.55312	4	181.65975
La1Pr4Zn3Co8O24	11.57062	5	172.42020
La1Y4Mn8Zn3O24	11.66563	5	121.20444
La5Zn3Fe8O24	11.75313	4	176.19335

Application

Assume that we are working on the stability of perovskite materials. Data contains **atomic volume, material composition,** and **number of elements.**

Material composition	Number of elements
Mg8Fe8O24	3
La1Y4Zn3Ni8O24	5
La1Y4Zn3Co8O24	5
La5Zn3Ni8O24	4
La1Pr4Zn3Ni8O24	5
La1Y4Zn3Fe8O24	5
La5Zn3Co8O24	4
La1Pr4Zn3Co8O24	5
La1Y4Mn8Zn3O24	5
La5Zn3Fe8O24	4

Type (number of elements)	Frequency
3-element material	1
4-element material	3
5-element material	6

Relative Frequency Distribution

The **class relative frequency** (CRF) is the class frequency (CF) divided by the total number of observations (n) in the data:

$$CRF = CF / n$$

The **class percentage** (CP) is the class relative frequency multiplied by 100:

$$CP = CRF \times 100$$

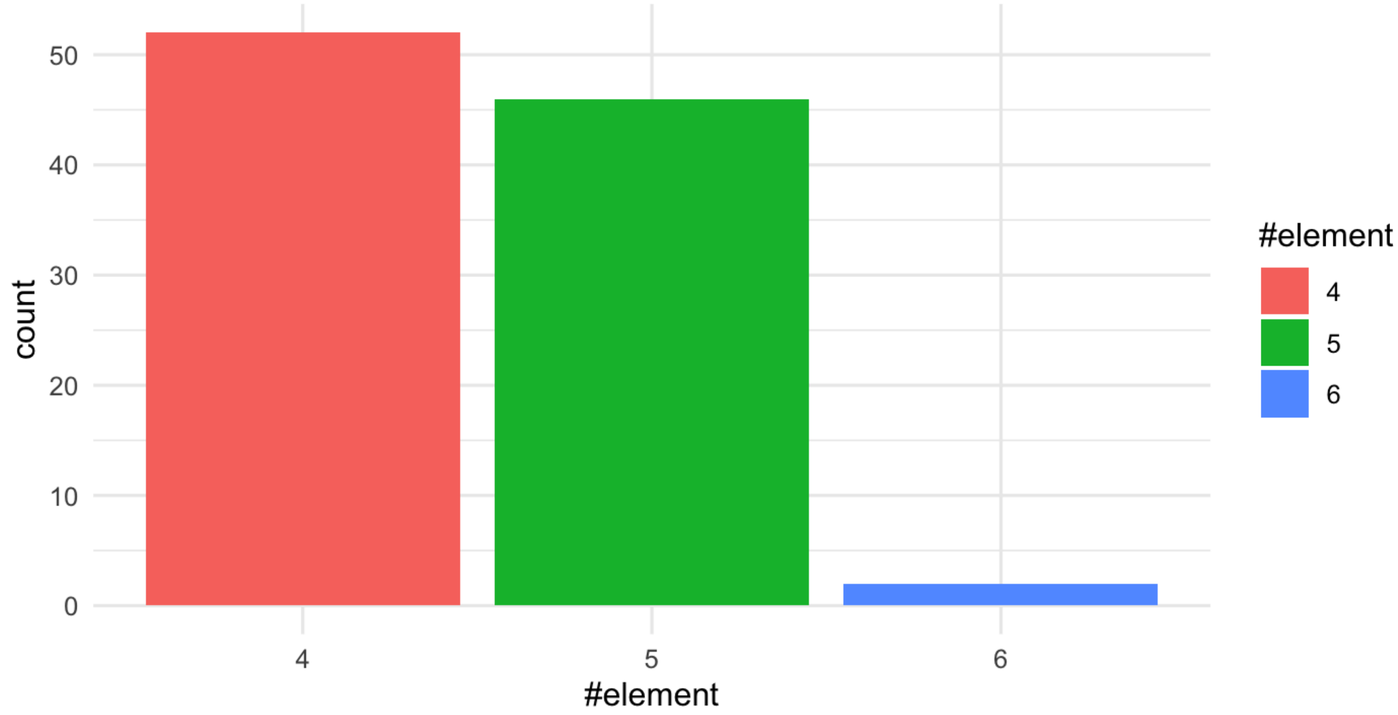
Application

Assume that we are working on the stability of perovskite materials. Data contains **atomic volume, material composition,** and **number of elements.**

Type (number of elements)	**Frequency**	**Relative frequency**	**Percent**
3-element material	1	1 / 10 = 0.1	%10
4-element material	3	3 / 10 = 0.3	%30
5-element material	6	6 / 10 = 0.6	%60

Bar plot

The categories (classes) of the qualitative data are represented by bars, where the height of each bar is either the class frequency, class relative frequency, or class percentage.



Pie chart

The categories of the qualitative data are represented by slices of a pie.

The size of each slice is proportional to the class relative frequency.

More plots?

You can **check the following websites out** to see more type of plots:

- https://r-graph-gallery.com/
- https://datavizproject.com/
- https://www.data-to-viz.com/
- https://datavizcatalogue.com/

Course materials

You can download the notes and codes from:

https://github.com/mcavs/ESTUMatse_2022Fall_EngineeringStatistics

Contact

Do not hesitate to contact me on:

https://twitter.com/mustafa__cavus

https://www.linkedin.com/in/mustafacavusphd/

mustafacavus@eskisehir.edu.tr