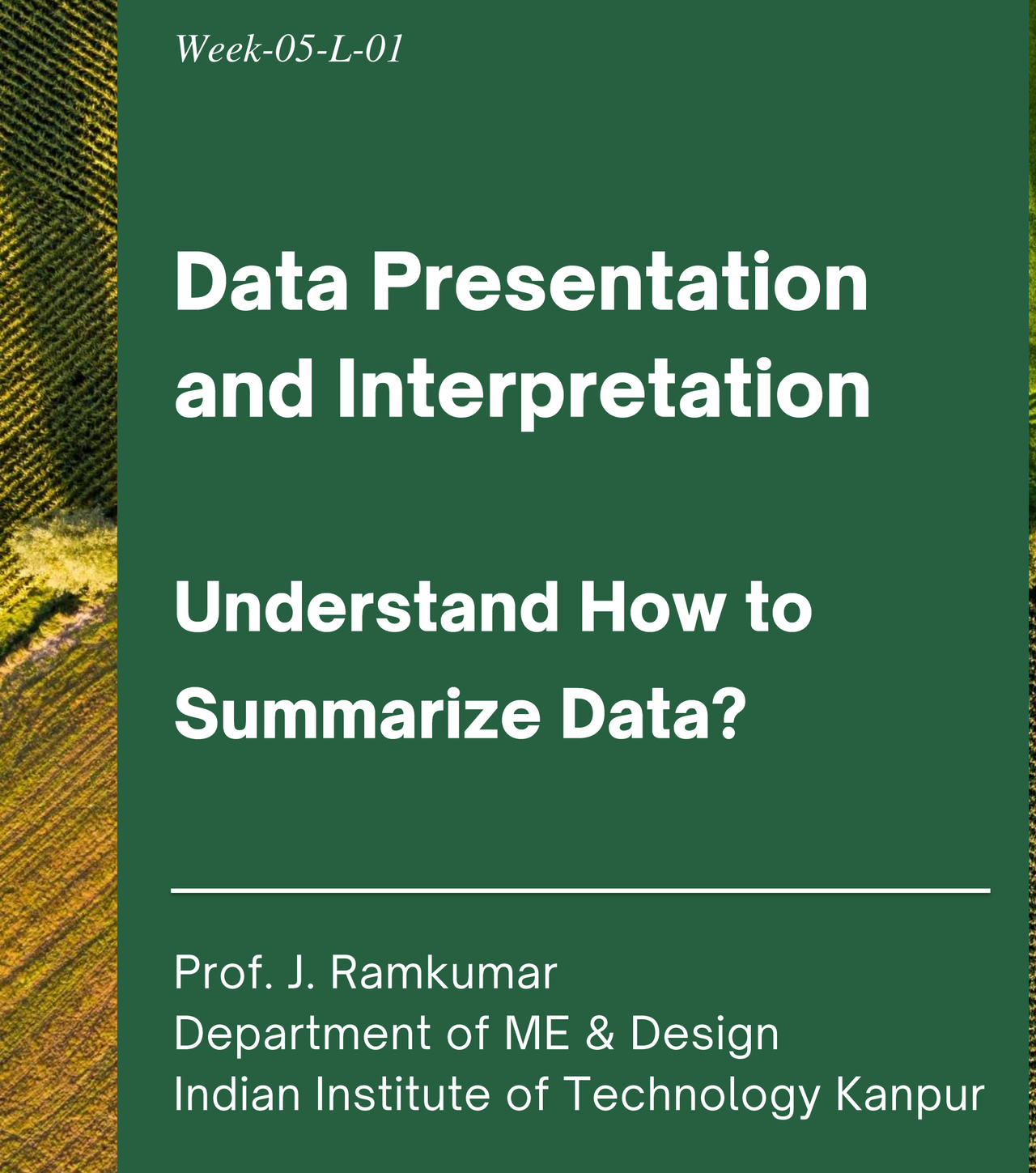
Week-05-L-01

Data Presentation and Interpretation

Understand How to Summarize Data?

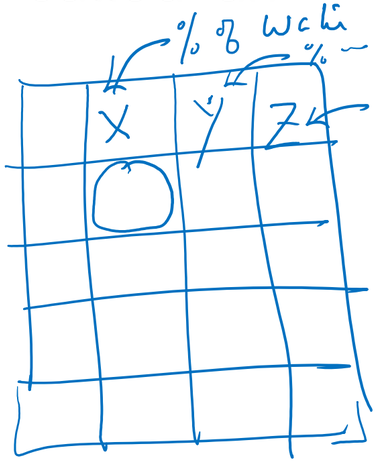
Prof. J. Ramkumar
Department of ME & Design
Indian Institute of Technology Kanpur



Important terms

Observation

- A dataset contains information about 'individuals'. Each 'individual' is called an 'observation' or 'case'.
- In most datasets, each row contains information about an individual.



Variable

- Any characteristic of an individual (i.e., observation) is called a variable. Some variables, like gender and job title, simply place individuals into categories.
- Others, like height or number of registered voters, take on numerical values for which we can do arithmetic.

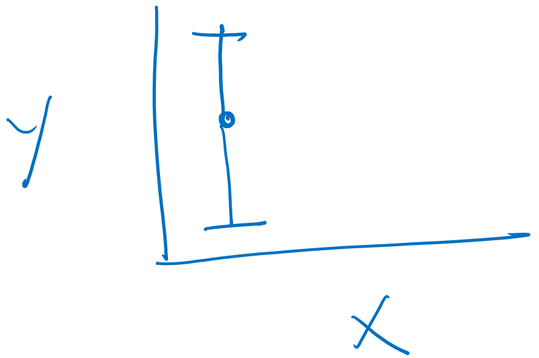
Describing and Summarizing

The two most useful ways of describing the distribution of data are:

1. **The typical:** This describes the center–or middle–of the data. This way of describing the center is also called a 'measure of central tendency'.

2. **The spread of the values around the center:** This describes how densely the data is distributed around the center. This is also called a 'measure of dispersion'.

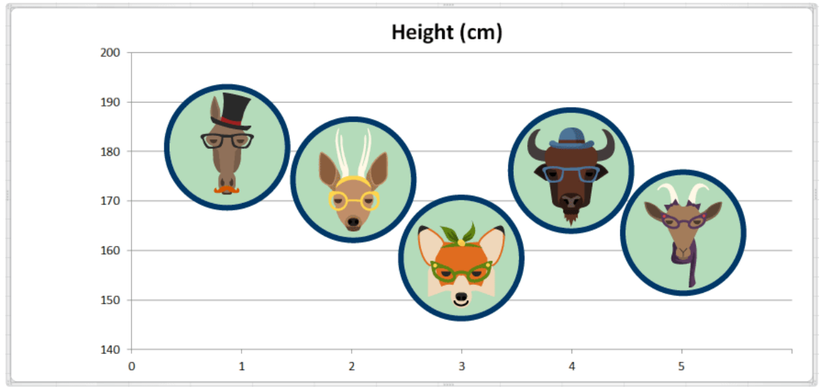
These two ways of describing the data are also referred to as descriptive statistics.

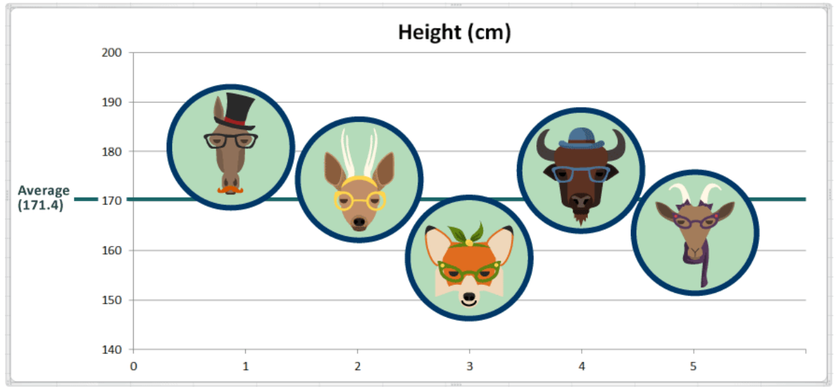


Central Tendencies (recalling)

- Mean
- Median
- Mode

	A	B
1	**Name**	**Height (cm)**
2	Harry the Horse	181
3	Dana the Deer	175
4	Fran the Fox	159
5	Bob the Buffalo	177
6	Gracie the Goat	165





Source: www.openelectiondata.net

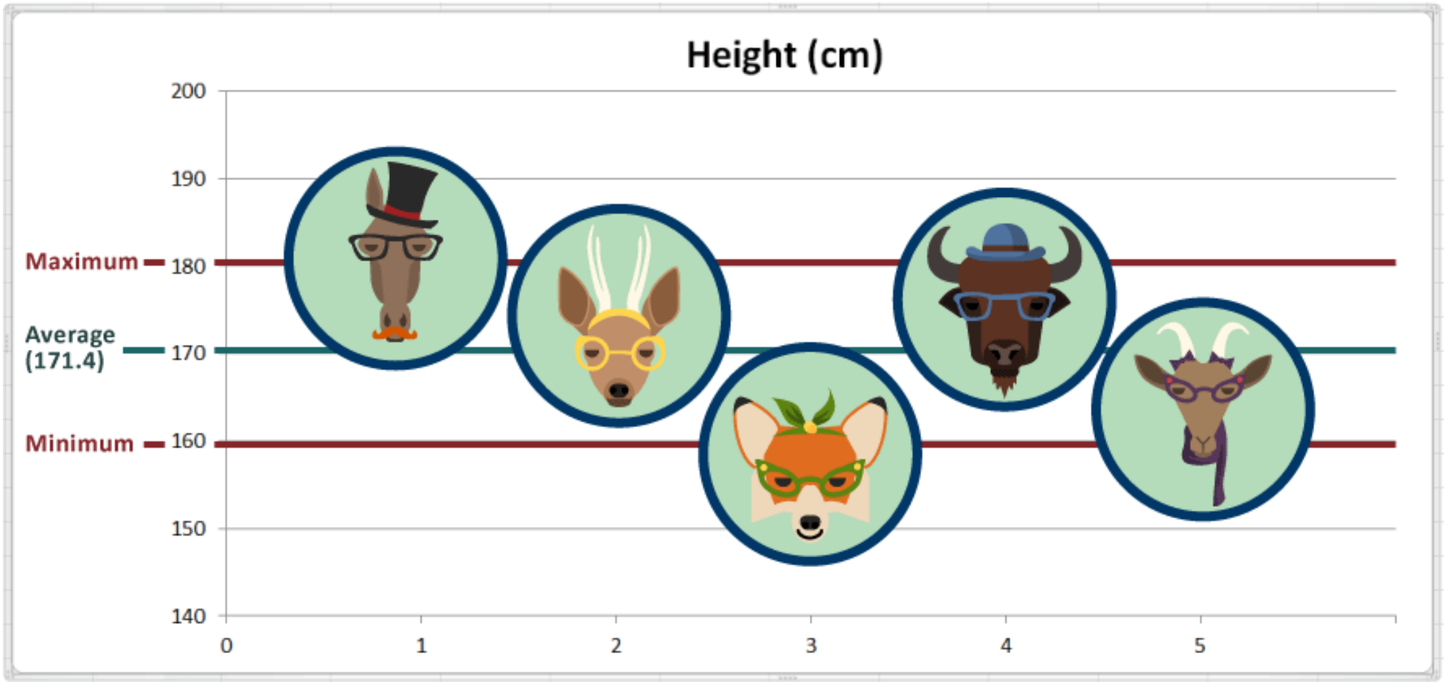
Measures of Dispersion (recalling)

- This is the difference between the largest and the smallest values.
- It is the distance between the extremes.
- Standard Deviation

23 ± 5 cm

$=$

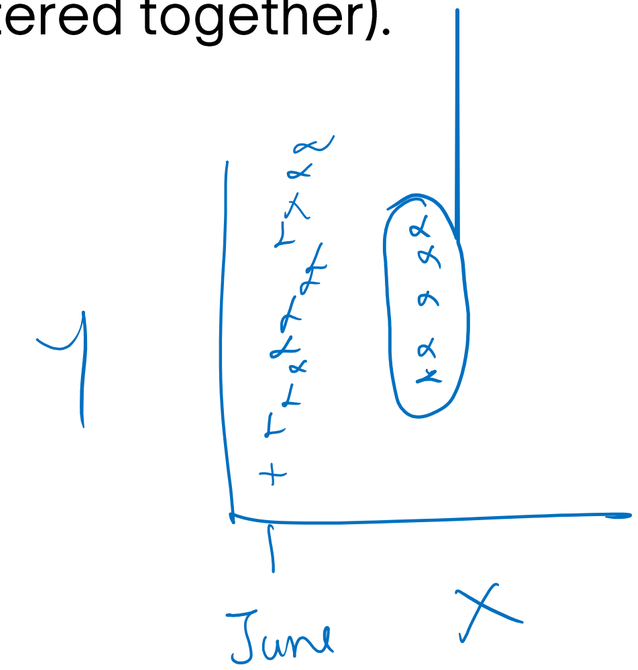
23 ± 20 cm ✗

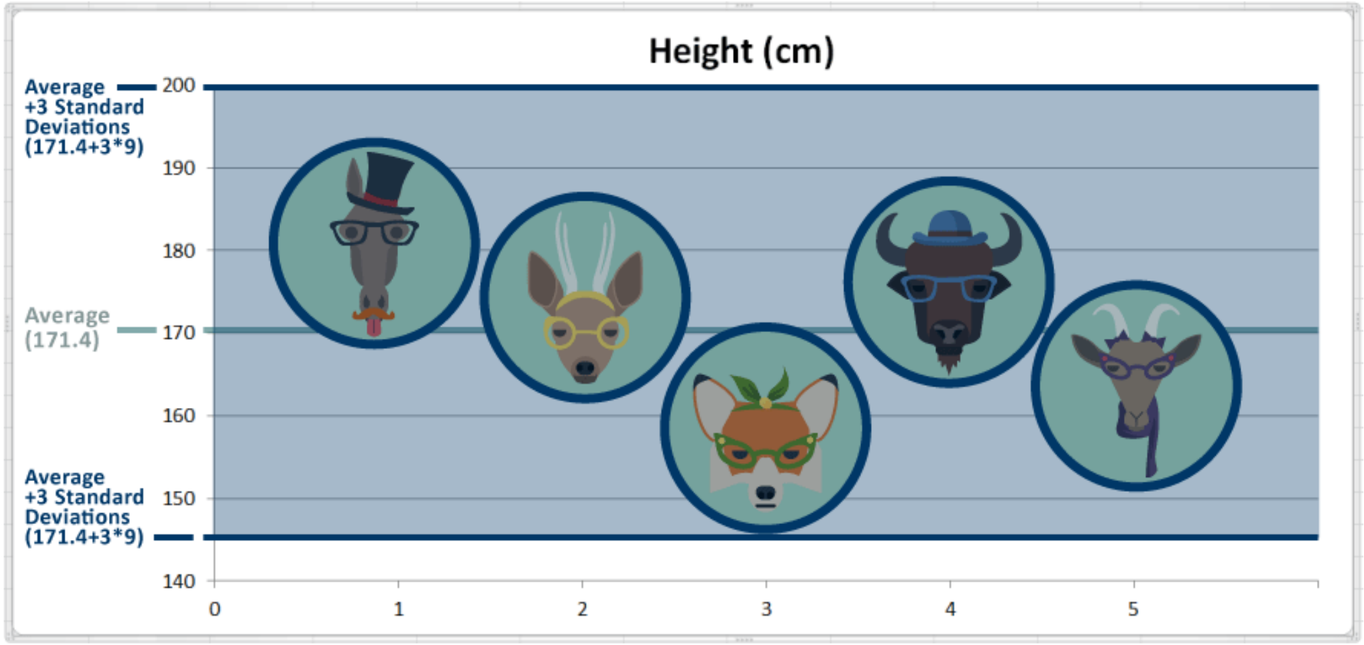


Source:
www.openelectiondata.net

Standard Deviation

- The standard deviation is like an 'index of variability,' because it is proportional to the scatter of the data.
- The standard deviation is larger for more diverse distributions (i.e., the data are widely scattered).
- The standard deviation is smaller for less diverse distributions (i.e., the data are clustered together).





Source: www.openelectiondata.net

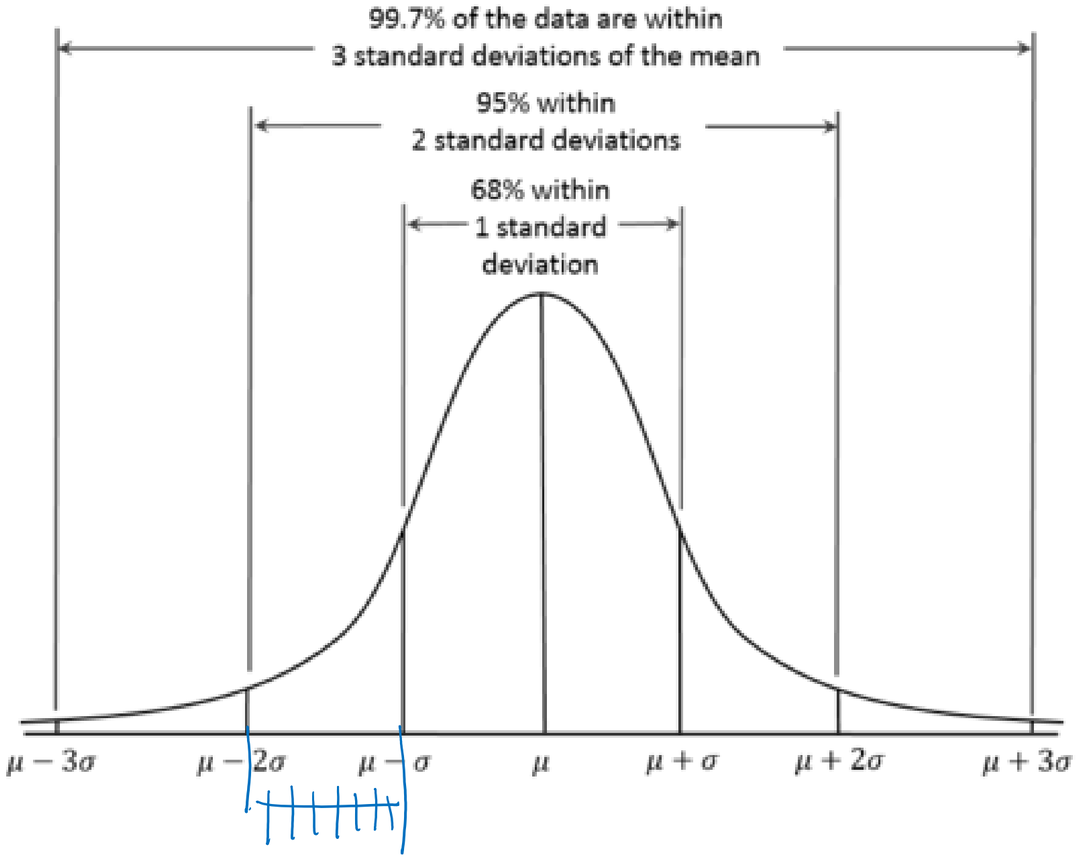
Standard Deviation

- The Population Standard Deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- The Sample Standard Deviation

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$



Source:
www.openelectiondata.net

Thank you