

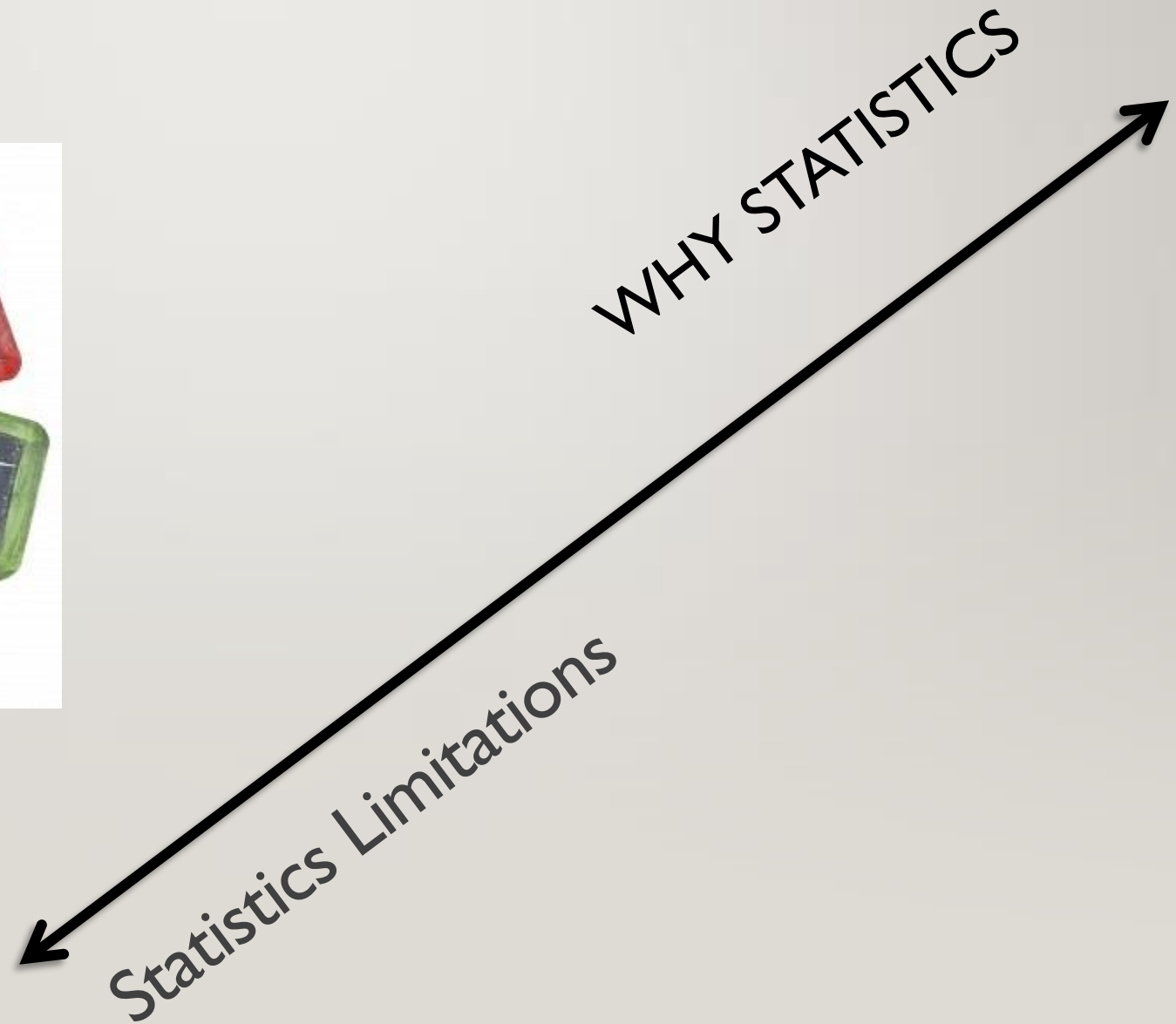
DSC Program 2020-2021



STATISTICS

FIRST SESSION

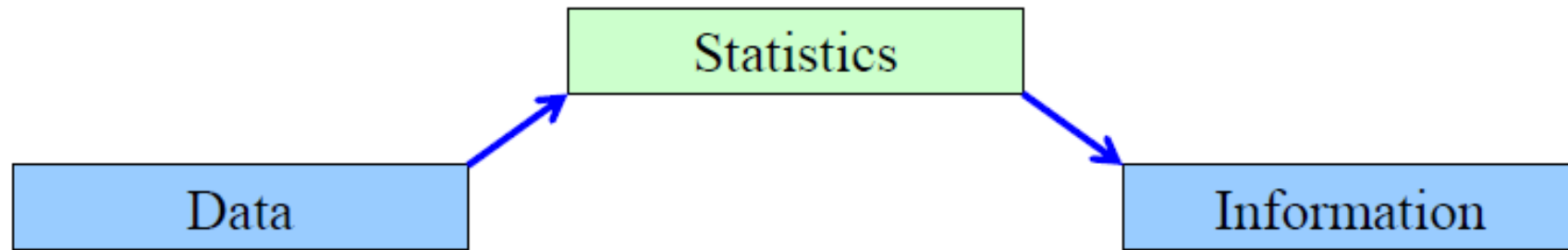
Statistics



“**STATISTICS** IS THE SCIENCE OF
COLLECTING ,ORGANIZING PRESENTING,
ANALYZING,AND INTERPRETING
NUMERICAL DATA”

“**STATISTICS** IS A WAY TO GET
INFORMATION FROM DATA.”

“Statistics is a way to get information from data”



Data: Facts, especially numerical facts, collected together for reference or information.

Information: Knowledge communicated concerning some particular fact.

Statistics is a *tool* for creating *new understanding* from a set of numbers.

Statistical Terms

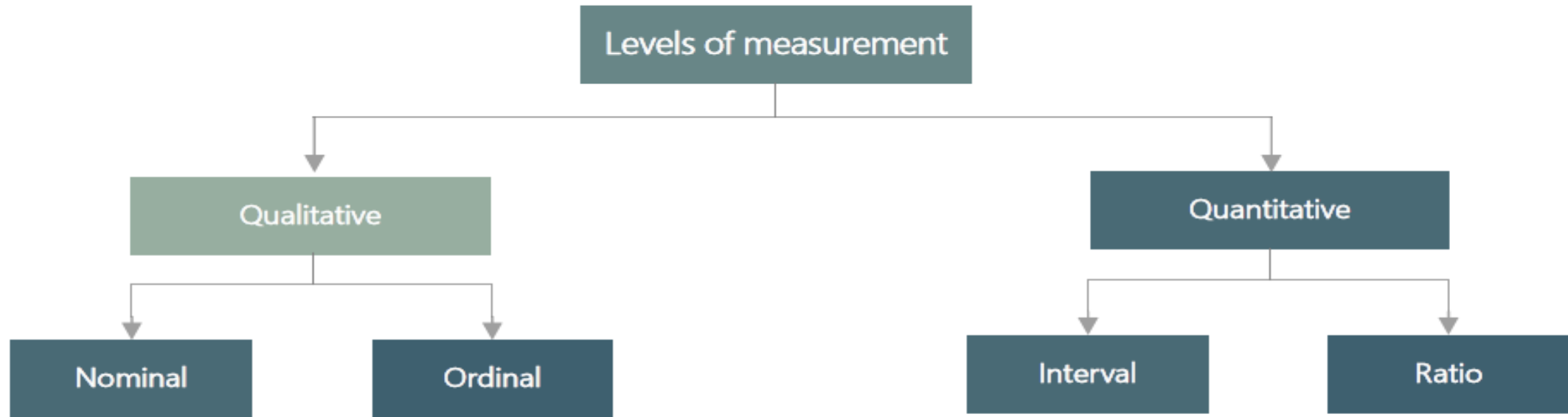
Variable



A variable is a feature characteristic of any member of a population differing in quality or quantity from another member.



Levels of measurement



There are two qualitative levels: nominal and ordinal. The nominal level represents categories that cannot be put in any order, while ordinal represents categories that **can** be ordered.

Examples:

Nominal: four seasons (winter, spring, summer, autumn)

Ordinal: rating your meal (disgusting, unappetizing, neutral, tasty, and delicious)

There are two quantitative levels: interval and ratio. They both represent "numbers", however, ratios **have a true zero**, while intervals don't.

Examples:

Interval: degrees Celsius and Fahrenheit

Ratio: degrees Kelvin, length

Types of data

```
graph TD; A[Types of data] --> B[Categorical]; A --> C[Numerical]; C --> D[Discrete]; C --> E[Continuous];
```

Categorical

Categorical data represents groups or categories.

Examples:

1. Car brands: Audi, BMW and Mercedes.
2. Answers to yes/no questions: yes and no

Numerical

Discrete

Continuous

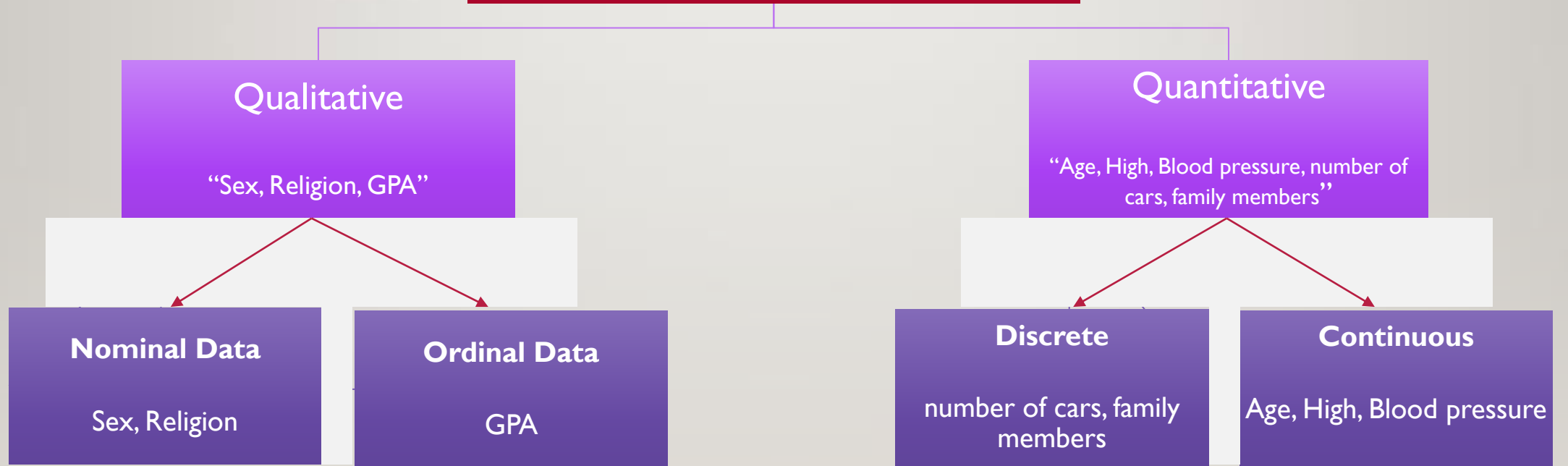
Numerical data represents numbers. It is divided into two groups: discrete and continuous. Discrete data can be usually counted in a finite matter, while continuous is infinite and impossible to count.

Examples:

Discrete: # children you want to have, SAT score

Continuous: weight, height

Statistical Data Type According to variable type



Statistical Terms

Population



A population is the group from which data is to be collected.

ALL



Population

Sample



A sample is a subset of a population.

SELECTED



Population



Sample

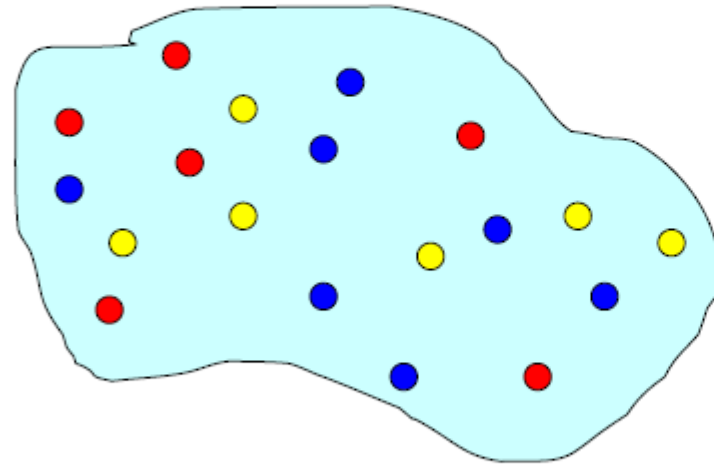
Parameter

A descriptive measure of a *population*.

Statistic

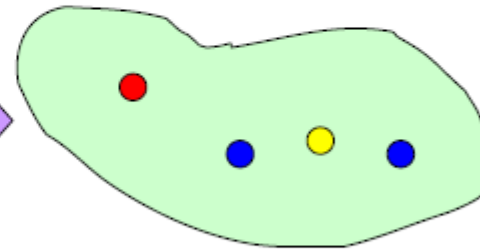
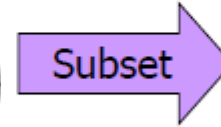
A descriptive measure of a *sample*.

Population



Parameter

Sample



Statistic

Populations have Parameters,

Samples have Statistics.

Measures of Central Tendency



Statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution.

DEFINITION

HOW TO COMPUTE

PROPERTIES

DEFINITIONS

Mean

- The sum of values divided by the number of data points.

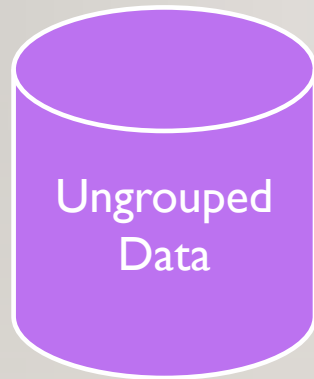
Median

- The middle value when the data arranged.

Mode

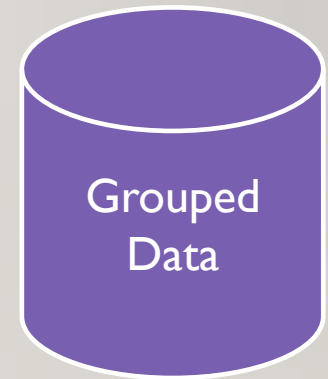
- The value that occurs most frequency

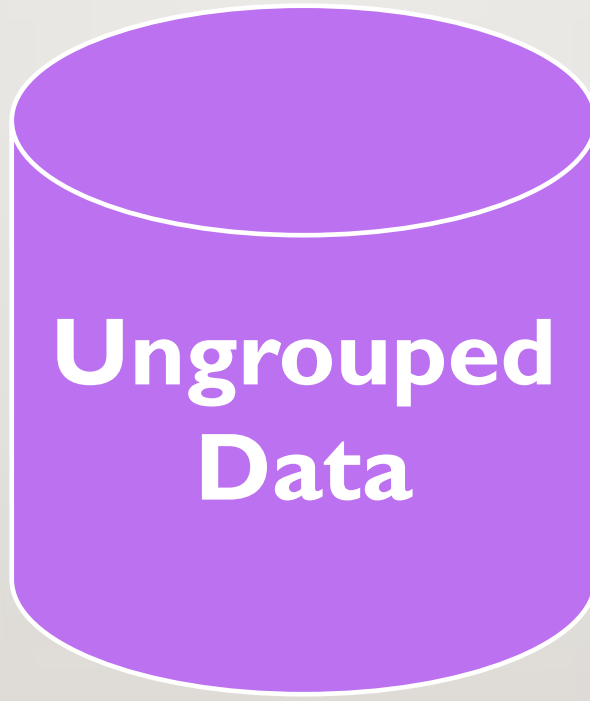
HOW TO COMPUTE



height
130
100
90
115
105
120
80
70

Height (in cm)	No of students
159-162	1
163-166	4
167-170	11
171-174	12
175-178	6
179-182	4
183-186	2





**Ungrouped
Data**

Mean

Ungrouped data

$$\text{Arithmetic Mean } \bar{x} = \frac{\sum x_i}{n}$$

	Height
1	130
2	100
3	90
4	115
5	105
6	120
7	80
8	70
Total	810

n=8

$$\text{mean} = 810/8 = 101.25$$

median

Ungrouped data

**Put the observation in ascending order(lowest first to highest last)

	Height
1	130
2	100
3	90
4	115
5	105
6	120
7	80
8	70

**1-Ascending
order**



n=8

	Height
1	70
2	80
3	90
4	100
5	105
6	115
7	120
8	130

When the number of observations (n) is **even**:

1. Find the value at position $\left(\frac{n}{2}\right)$

$$n/2 = 4 \quad v1 = 100$$

2. Find the value at position $\left(\frac{n}{2} + 1\right)$

$$(n/2) + 1 = 5 \quad v2 = 105$$

$$\text{median} = (v1 + v2) / 2 = (100 + 105) / 2 = 102.5$$

	Weight
1	62
2	55
3	48
4	90
5	52
6	60
7	40
8	50
9	60

**1-Ascending
order**



n=9

	Weight
1	40
2	48
3	50
4	52
5	55
6	60
7	60
8	62
9	90

When the number of observations (n) is **odd**:
the median is the value at position

$$\left(\frac{n+1}{2}\right)$$

$$(n+1)/2 = 5$$

$$\text{median} = 55$$

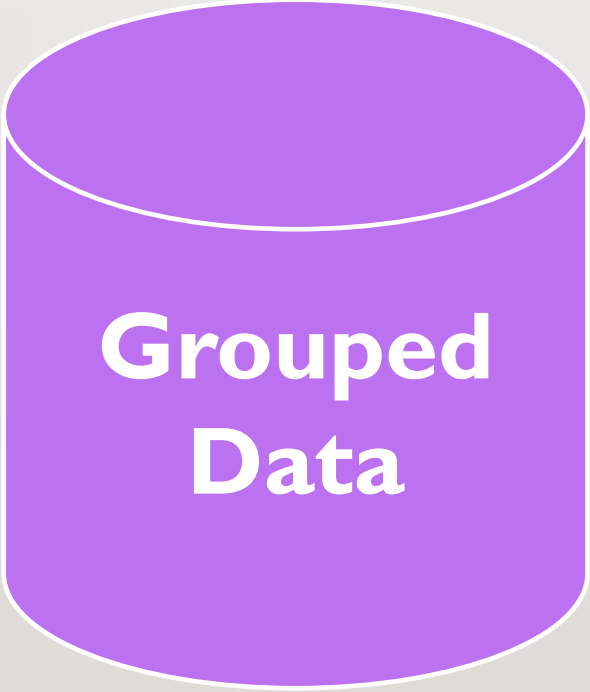
The value that occurs most frequency

Mode

Grouped data

	Height	Weight
1	130	62
2	100	55
3	90	48
4	115	90
5	105	52
6	120	60
7	80	40
8	70	50
9		60

height mode=	nan
weight mode=	60



**Grouped
Data**

Mean, median and mode

Grouped data

Class	frequency(f)
16-	4
24-	8
32-	9
40-	13
48-	10
56-	5
64-72	1
Total	50.00

Mean

Grouped data

$$\text{Arithmetic Mean} = \frac{\sum (f_i * x_i)}{\sum f_i}$$

Class	frequency(f)	Mid-Point(x)	f * x
16-	4	20	80
24-	8	28	224
32-	9	36	324
40-	13	44	572
48-	10	52	520
56-	5	60	300
64-72	1	68	68
Total	50.00		2088

$$\bar{X} = \frac{\sum fx}{\sum f} = \frac{2088}{50} = 41.76$$

Median

Grouped data

Class	frequency(f)
16-	4
24-	8
32-	9
40-	13
48-	10
56-	5
64-72	1
Total	50.00

1-Ascending cumulative frequency distribution table

Upper limits	CF
Less than 24	4
Less than 32	=4+8=12
Less than 40	=12+9=21
Less than 48	34
Less than 56	44
Less than 64	49
Less than 72	50

2-Find the position of median = $\frac{\Sigma f}{2} = \frac{50}{2} = 25$

3-Median class = (40-48)
Lower limit of median class (M_o)=40
Width of median class (l)=8
 $f_1 = 21$
 $f_2 = 34$

$$\begin{aligned} \text{median} &= M_o + \frac{\text{pos.} - f_1}{f_2 - f_1} * l \\ &= 40 + \frac{25 - 21}{34 - 21} * 8 = 42.46 \end{aligned}$$

Mode

Grouped data

$l=8$ equal width

Class	frequency(f)
16-	4
24-	8
32-	9
40-	13
48-	10
56-	5
64-72	1
Total	50.00

$M_o=40$

$$\Delta_1 = 13 - 9 = 4$$

$$\Delta_2 = 13 - 10 = 3$$

$$mode = M_o + \frac{\Delta_1}{\Delta_1 + \Delta_2} * l$$

$$mode = 40 + \frac{4}{4 + 3} * 8$$

$$mode = 42.46$$

Mode
Grouped data
unequal width

Class	frequency (f)	The length of the class	modified frequency
0-	3	5	3/5=0.60
5-	8	7	8/7=1.14
12-	16	8	2.00
20-	11	10	1.10
30-	7	15	0.47
45-50	5	5	1.00
total	50		

$M_o = 12$

$l = 8$

$$\Delta_1 = 2 - 1.143 = 0.857$$

$$\Delta_2 = 2 - 1.1 = 0.9$$

$$\begin{aligned} \text{mode} &= M_o + \frac{\Delta_1}{\Delta_1 + \Delta_2} * l \\ \text{mode} &= 12 + \frac{0.857}{0.857 + 0.9} * 8 \\ \text{mode} &= 15.9 \end{aligned}$$

Properties

Which
measure is
best

Mean

unique

Simplicity

Affected by outliers

$$\sum (x_i - \bar{x}) = 0$$

Median

unique

Simplicity

Not affected

Preferred in case of ranked data

Mode

Not unique

Simplicity

There is no best, but using only one is definitely the worst!!!

Measures of Dispersion

Dispersion refers to measures of how spread out our data is. Typically they're statistics for which values near zero signify *not spread out at all* and for which large values (whatever that means) signify *very spread out*.

DEFINITION

HOW TO COMPUTE

PROPERTIES

RANGE & IQR

Five Number Summary

MAXIMUM

THIRD QUARTILE

SECOND QUARTILE (MEDIAN)

FIRST QUARTILE

MINIMUM

RANGE & IQR

Finding the 5 number summary

8,1,3,5,10,2,3

1- data sort

1, 2, 3, 3, 5, 8, 10

2- Range= max - min
= 10 - 1 = 9

3- Interquartile Range
= Q3 - Q1 = 8 - 2 = 6

MINIMUM	Q1	Q2 (MEDIAN)	Q3	MAXIMUM
1	2	3	8	10

VARIANCE



Variance measures the dispersion of a set of data points around their mean

VARIANCE

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



population
variance

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$



sample
variance

STANDARD DEVIATION FORMULAS

$$\sigma = \sqrt{\sigma^2}$$

population standard
deviation

sample standard
deviation

$$S = \sqrt{S^2}$$

COEFFICIENT OF VARIATION (CV)

relative standard
deviation/

standard deviation

mean

365√DataScience

COEFFICIENT OF VARIATION (CV)

$$C_v = \frac{\sigma}{\mu}$$

Population formula

Sample formula

$$\hat{C}_v = \frac{s}{\bar{x}}$$

365√DataScience



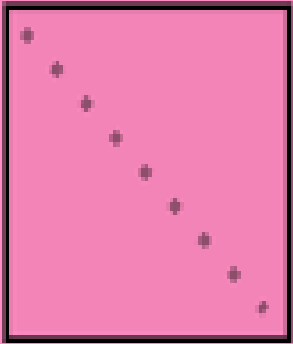
Standard deviation is the most common measure of variability for a SINGLE DATASET

Comparing TWO OR MORE datasets

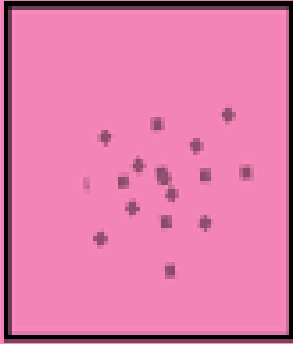


Covariance VS Correlation

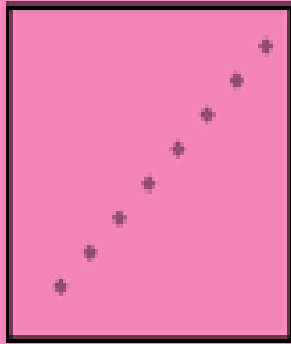
COVARIANCE



Large Negative
Covariance

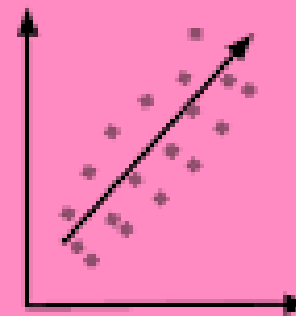


Nearly Zero
Covariance

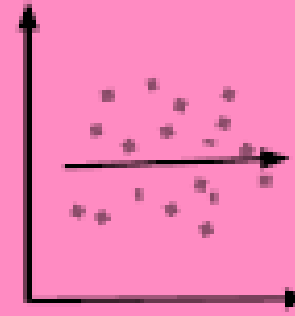


Large Positive
Covariance

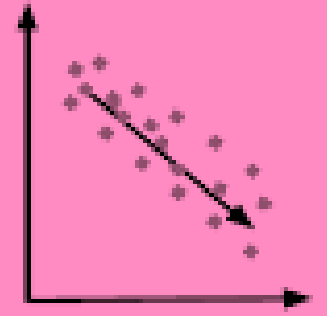
CORRELATION



Positive
Correlation



Zero
Correlation



Negative
Correlation

Skewness

Positive (right)

Dataset 1	Interval	Frequency
1	0 to 1	4
1	1 to 2	6
1	2 to 3	4
1	3 to 4	2
2	4 to 5	2
2	5 to 6	0
2	6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00

Zero (no skew)

Dataset 2	Interval	Frequency
1	0 to 1	2
1	1 to 2	2
2	2 to 3	3
2	3 to 4	5
3	4 to 5	3
3	5 to 6	2
3	6 to 7	2

Mean	Median	Mode
4.00	4.00	4.00

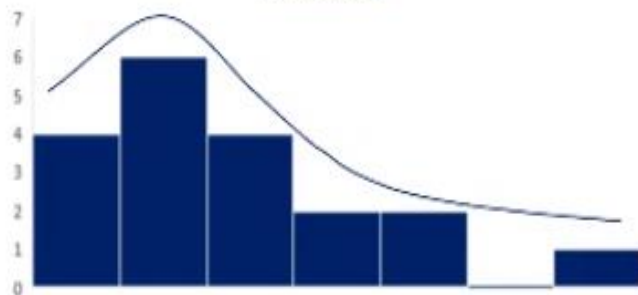
Negative (left)

Dataset 3	Interval	Frequency
1	0 to 1	1
2	1 to 2	1
3	2 to 3	2
3	3 to 4	3
4	4 to 5	4
4	5 to 6	6
4	6 to 7	3

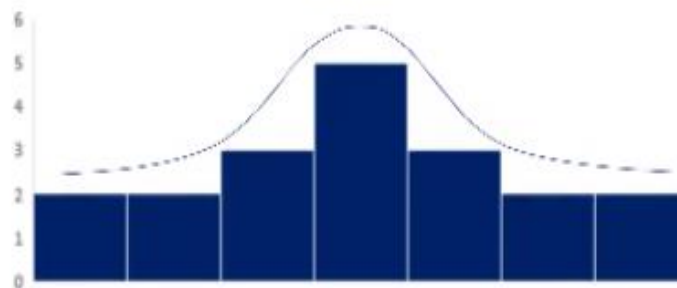
Mean	Median	Mode
4.90	5.00	6.00

mean < median

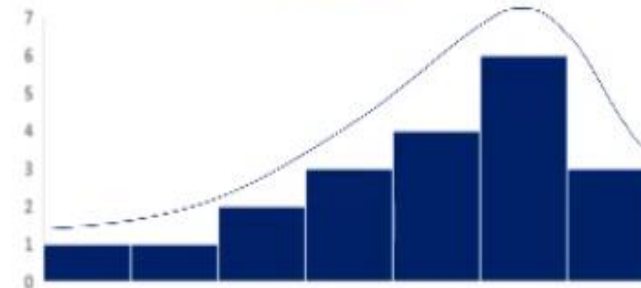
Positive skew



Zero skew



Negative skew



Skewness

The coefficient of Skewness is a measure for the degree of symmetry in the variable distribution.



Negatively skewed distribution
or Skewed to the left
Skewness < 0



Normal distribution
Symmetrical
Skewness $= 0$



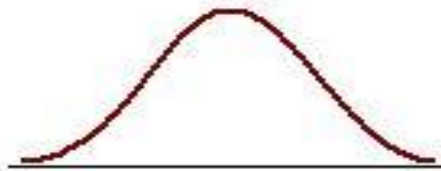
Positively skewed distribution
or Skewed to the right
Skewness > 0

Kurtosis

The coefficient of Kurtosis is a measure for the degree of peakedness/flatness in the variable distribution.



Platykurtic distribution
Low degree of peakedness
Kurtosis < 0



Normal distribution
Mesokurtic distribution
Kurtosis $= 0$



Leptokurtic distribution
High degree of peakedness
Kurtosis > 0

Feature Scaling

Standardization Vs Normalization

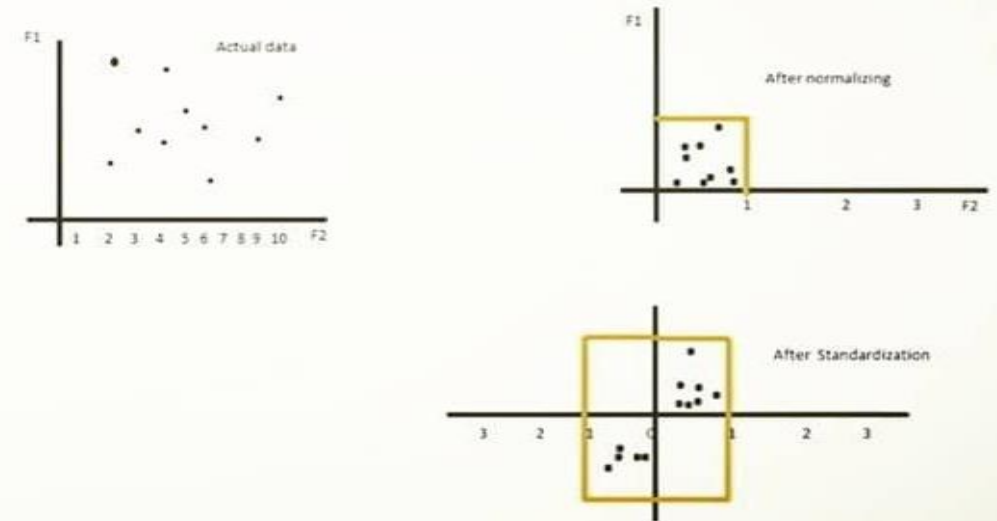
Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

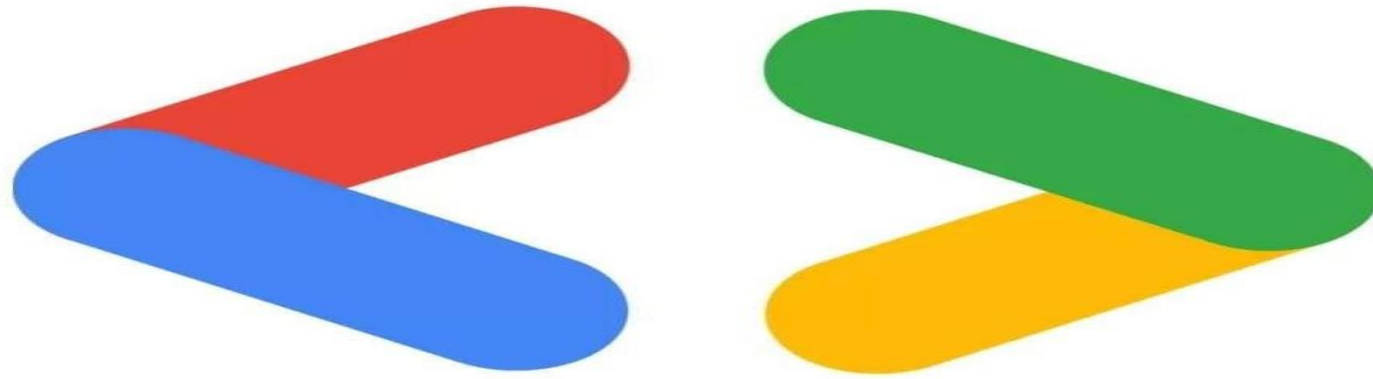
$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

$$X_{changed} = \frac{X - \mu}{\sigma}$$

For most applications standardization is recommended.





Developer Student Clubs

Al-Azhar University

DSCOn Track

Thank you