STATISTISTICAL DATA ANALYSIS

TEACHERS

Manuella Lech Cantuaria
Assistant Professor, PhD
The Maersk Mc-Kinney Moller Institute
Applied AI and Data Science
mlca@mmmi.sdu.dk

Victoria Blanes-Vidal
Associate Professor, PhD
The Maersk Mc-Kinney Moller Institute
Applied AI and Data Science
vbv@mmmi.sdu.dk

Research line: Data science and machine learning applied to health data and epidemiology

ABOUT THE COURSE

- **Learning objectives - Knowledge**
  - explain relevant data types and their representation for statistical analysis
  - explain probabilities and random variables
  - explain distributions of random variables
  - explain inference and hypothesis testing
  - explain how data may be collected from experiments involving randomness
- **Learning objectives - Skills**
  - choose an appropriate experimental design in respect to a given task
  - perform statistical analyzes on data collected
  - use a statistical tool for analysis and visualization of data
- **Learning objectives - Competences**
  - use statistical methods and tools to interpret experimental data

RECOMMENDED LITERATURE

- OpenIntro Statistics (David Diez, Mine Cetinkaya-Rundel, Christopher Barr).
  - It can be downloaded for free here: https://www.openintro.org/book/os/
- Other recommended books:
  - Applied statistics and probability for engineers / Douglas C. Montgomery, George C. Runger —3rd ed. ISBN 0-471-20454-4
  - A Handbook of Statistical Analyses Using R / Brian S. Everitt, Torsten Hothorn , ISBN 1420079336

COURSE ORGANIZATION

- Thursdays 10-12: Main lecture (Manuella or Victoria)
- Thursdays 12-14: Exercise time (independent work on exercises with help of instructors). There will be one instructor responsible per class.
  - Instructors are:
    - Henrik Dyrberg Egemose
    - Sofie Ørnfeldt Nedergaard
    - Christian Maretti Wann Bengtsen
    - Jonathan Wanjau Leegaard Riis
    - Lasse Schier Christiansen

LECTURE PLANNING

Lesson	Week	Date	TOPICS	Teacher (planned)
1	36	9/Sep	Introduction to the course Descriptive statistics – part I	Manuella
2	37	16/sep	Descriptive statistics – part II	Manuella
3	38	23/Sep	Probability distributions	Manuella
4	39	30/Sep	Hypothesis testing (one sample)	Victoria
5	40	7/Oct	Hypothesis testing (two samples)	Victoria
-	41	14/Oct	NO CLASS	
-	42	21/Oct	NO CLASS (Autum holidays)	
6	43	28/Oct	ANOVA one-way	Victoria
7	44	4/Nov	R class (hypothesis testing + ANOVA)	Manuella
8	45	11/Nov	ANOVA two-way Notions of experimental design	Victoria
9	46	18/Nov	Regression analysis	Victoria
10	47	25/Nov	Multiple regression	Manuella
11	48	2/Dec	Logistic regression	Manuella
12	49	9/Dec	Recap of statistical concepts, questions’ time, etc	Both

EXAM

- Multiple choice exam in January
  - Questions will involve concepts' understanding and calculations for problem solving
  - 120 minutes
  - Probably beginning of january – dates will come later
- Reexam in February

Chapter 1:
Descriptive Statistics

Manuella Lech Cantuaria

Assistant Professor, PhD

The Maersk Mc-Kinney Moller Institute

Applied AI and Data Science

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Data organization and histograms

1.4. Measures of:

- Central Tendency (Location)
- Variation (Dispersion)
- Position

1.5. Data representation: frequency distributions and graphs

1.6. Shapes of frequency distributions: Skewness and kurtosis

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential

1.2. Variables and Types of Data

1.3. Data organization and histograms

1.4. Measures of:

Central Tendency (Location)

Variation (Dispersion)

Position

1.5. Data representation: frequency distributions and graphs

1.6. Shapes of frequency distributions: Skewness and kurtosis

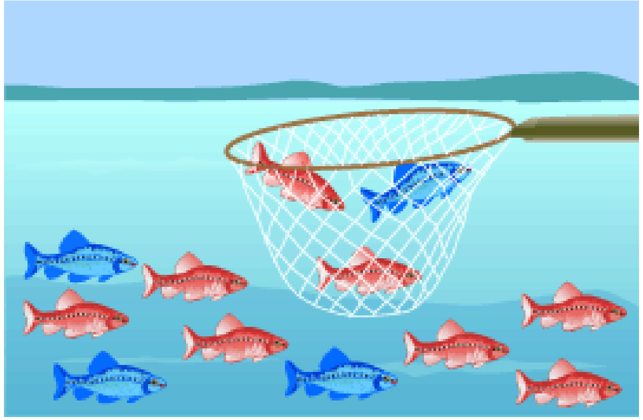
1-1 Statistics

- **Statistics** is the science of conducting studies to

  collect,
  organize,
  summarize,
  analyze, and
  draw conclusions from data.

1-1 Statistics

- A **variable** is a characteristic or attribute that can assume different values.
- The values that a variable assumes are called **data**.
- A **population** consists of all subjects (human or otherwise) that we want to study.
- A **sample** is a subset of the population.



1-1 Statistics

Population vs. Sample

- It is important to know, whether all data (the population) or only a subset (a sample) are known.

Sample	Population
A selection of 1000 inhabitants of a town	All inhabitants of a town
560 measurements of copper in the soil of a field	*Not possible*

1-1 Statistics

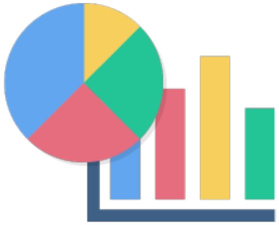
Descriptive Statistics

Used to describe the **sample** data

Tables

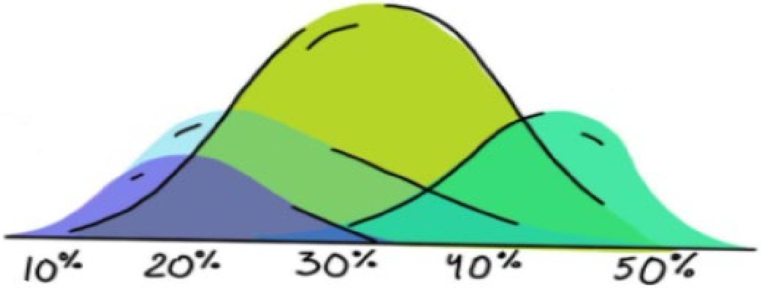
- Tables are extremely useful to summarize data upon conclusions are based.
- It uses a minimum of space to communicate a large amount of information.

Graphs



- More visual than tables
- Often preferred to show variable’s trends, better understand data distribution and variability and compare groups.

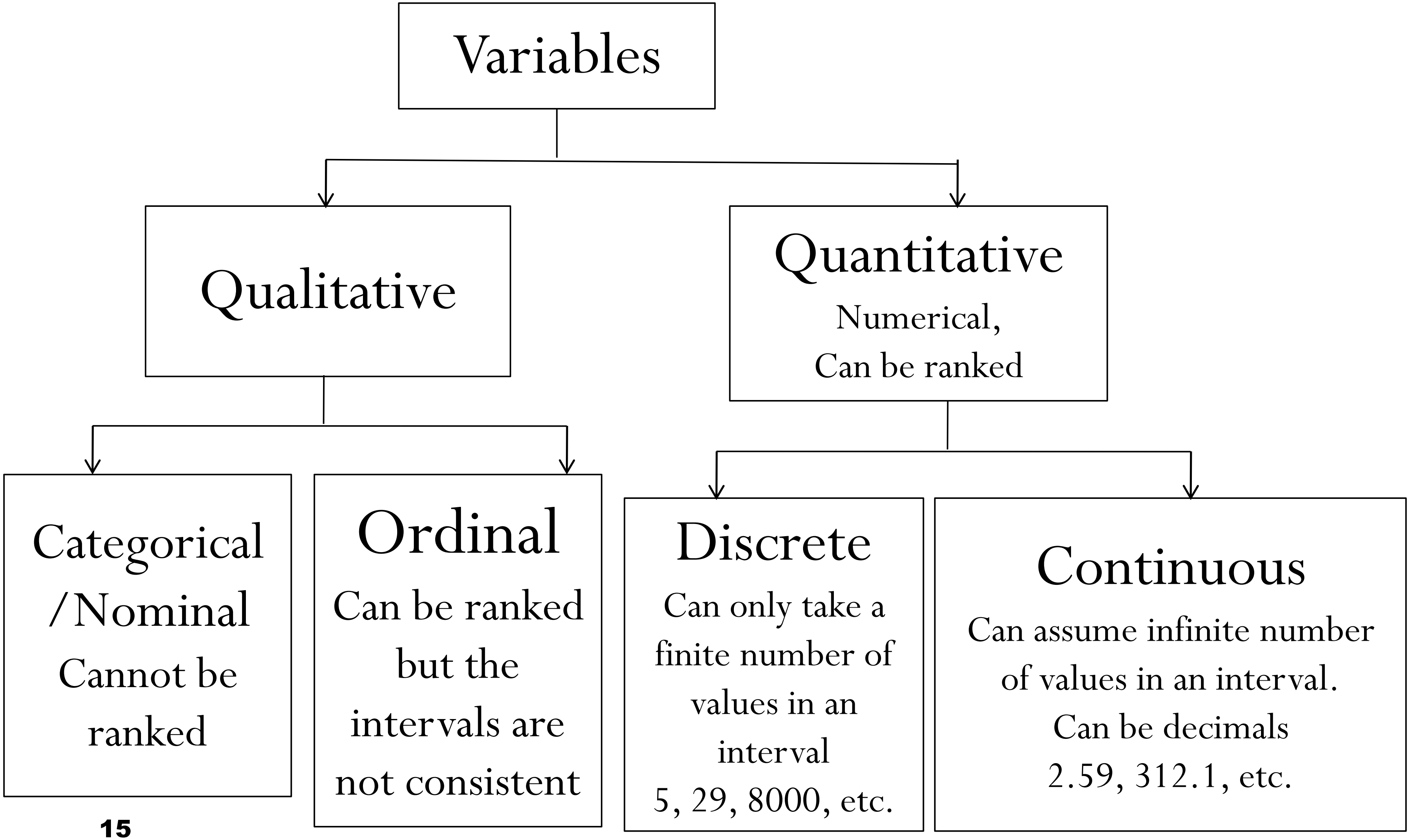
Inferential Statistics



Uses the **sample** data to draw conclusions about a **population**

1-2 Types of Variables and Data

- Variables can be classified as:



Variables

Qualitative

Quantitative
Numerical,
Can be ranked

Categorical /Nominal
Cannot be ranked

Ordinal
Can be ranked but the intervals are not consistent

Discrete
Can only take a finite number of values in an interval
5, 29, 8000, etc.

Continuous
Can assume infinite number of values in an interval.
Can be decimals
2.59, 312.1, etc.

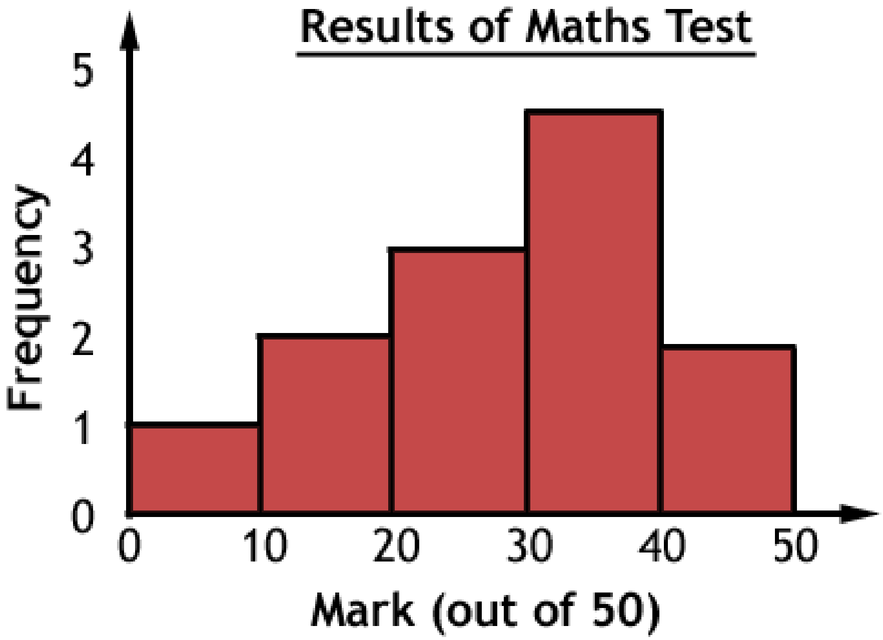
1-3 Data organization and histograms

- When conducting a statistical study, the researcher must gather data for the particular variable under study.

- To describe situations (descriptive statistics) or draw conclusions and make inferences about populations (inferential statistics), the researcher must organize the data in some meaningful way.

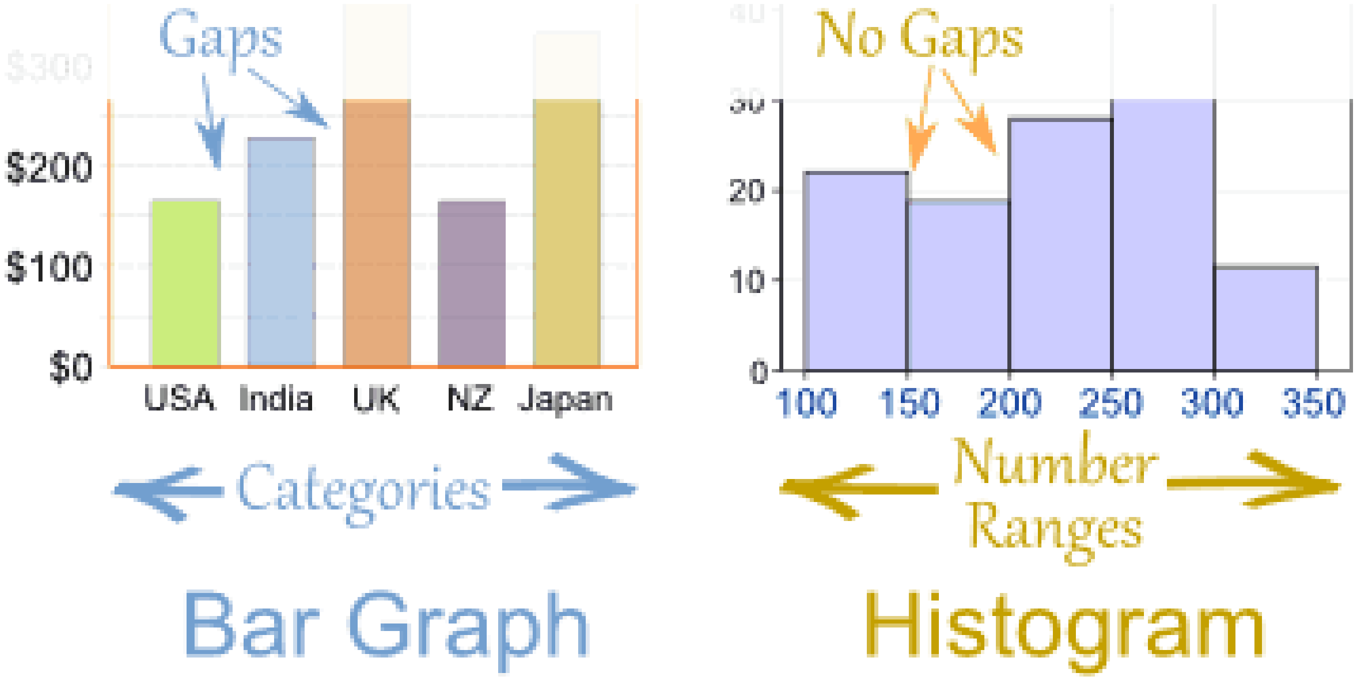
Histogram

The ***histogram*** is a graph that displays the data by using vertical bars of various heights to represent the frequencies of the classes.



The height of each bar represents the percentage (or counts) of data values in the interval

Bar chart and histograms



- For the description of the variability
- Divide range of possible measurements into a number of groups
- Count observations in each group

Building a Histogram

- **Daily low temperatures** recorded in a town (01/18-01/31, 2005, °F)

Jan. 18 – 11

Jan. 19 – 11

Jan. 20 – 25

Jan. 21 – 29

Jan. 22 – 27

Jan. 23 – 14

Jan. 24 – 11

Jan. 25 – 25

Jan. 26 – 33

Jan. 27 – 22

Jan. 28 – 18

Jan. 29 – 19

Jan. 30 – 30

Jan. 31 – 27

Building a Histogram

- **(1) Develop an ungrouped frequency table**

→ Data (minimum measured temperature: T_{min} (F)):

11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33

→

11	3
14	1
18	1
19	1
22	1
25	2
27	2
29	1
30	1
33	1

Building a Histogram

- **2. Construct a grouped frequency table**

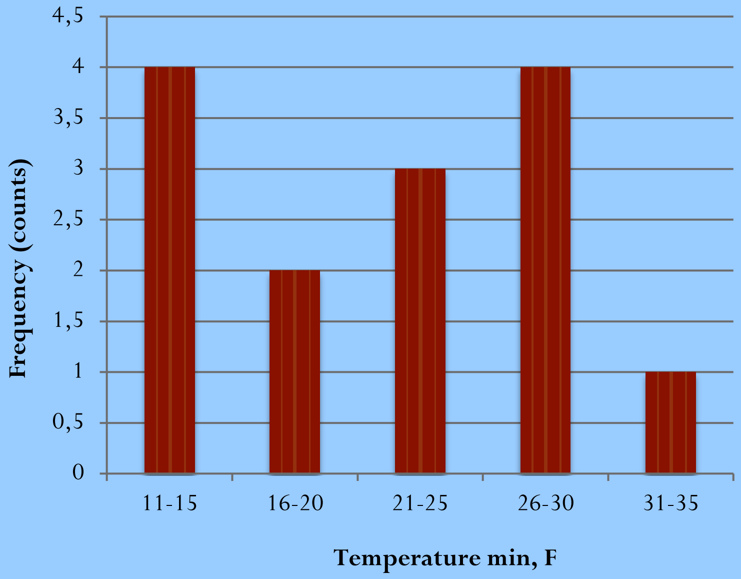
  → Select a set of classes

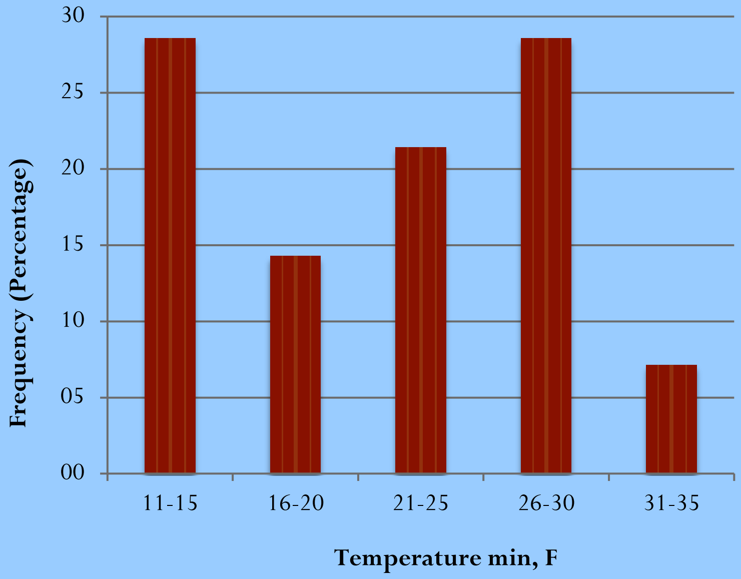
  →

11-15	4
16-20	2
21-25	3
26-30	4
31-35	1

Building a Histogram

- **3. Plot the frequencies of each class**





1-4 Measures of Central Tendency (location), Variation (dispersion) and Position

- The data distribution can be described with four characteristics:
  - Measures of location
  - Measures of dispersion
  - Measures of position
  - Skewness and Kurtosis

Measures of Central Tendency (Location)

What Do We Mean By **Average**?

- Mean
- Median
- Mode
- Midrange

Mean

- The **mean** is the division of the sum of the values and the total number of values.
- The symbol $\overline{X}$ is used for sample mean.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X}{n}$$

- For a population, the Greek letter μ (mu) is used for the mean.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum X}{N}$$

Note: General Rounding Rule

The basic rounding rule is that rounding should not be done until the final answer is calculated.

Median

- The **median** is the midpoint of the data array.
- How to calculate the median:
  - Sort in ascending order.
  - Select the middle value.
- The median will be one of the data values if there is an odd number of values.
- The median will be the average of two data values if there is an even number of values.

Mode

- The **mode** is the value that occurs most often in a data set.
- It is sometimes said to be the most typical case.
- There may be no mode, one mode (unimodal), two modes (bimodal), or many modes (multimodal).

Midrange

- The **midrange** is the average of the lowest and highest values in a data set.

$$MR = \frac{Lowest + Highest}{2}$$

An Example Data Set

- **Daily low temperatures** recorded in a town (01/18-01/31, 2005, °F)

Jan. 18 – 11
Jan. 19 – 11
Jan. 20 – 25
Jan. 21 – 29
Jan. 22 – 27
Jan. 23 – 14
Jan. 24 – 11
Jan. 25 – 25
Jan. 26 – 33
Jan. 27 – 22
Jan. 28 – 18
Jan. 29 – 19
Jan. 30 – 30
Jan. 31 – 27

- For these 14 values, we will calculate all four measures of central tendency - the **mean**, **median**, **mode**, and **midrange**

Mean

- **Mean** –Most commonly used measure of central tendency
- **Procedures**
- (1) **Sum** all the values in the data set
- (2) **Divide** the sum by the number of values in the data set
- Watch for **outliers**

An **outlier** is an observation point that is very distant from other observations

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median

- **Median** - **1/2** of the values **are above** it & **1/2 below**
- (1) **Sort** the data in **ascending** order
- (2) **Find** the value with an **equal number** of values above and below it
- (3) **Odd** number of observations → $[(n-1)/2]+1$ value from the lowest
- (4) **Even** number of observations → average $(n/2)$ and $[(n/2)+1]$ values

Mode

- **Mode** – This is the most frequently occurring value in the distribution
- (1) **Sort** the data in **ascending** order
- (2) **Count** the **instances** of each value
- (3) **Find** the value that has the **most** occurrences
- If more than one value occurs an **equal number** of times and these exceed all other counts, we have **multiple** modes
- Use the mode for **multi-modal** data

Midrange

- (1) **Sort** the data in ascending order:
- (2) **Select** the lowest and highest values:
- (3) **Find** the mean of those two values

Properties of the Mean

- Uses all data values.
- Unique, usually not one of the data values
- Affected by extremely high or low values, called outliers

Properties of the Median

- Affected less than the mean by extremely high or extremely low values.

Properties of the Mode

- Easy to compute.
- Can be used with nominal data
- May not exist

Properties of the Midrange

- Easy to compute.
- Affected by extremely high or low values in a data set

Questions?

Now it is time for you to practice at the exercise's class!
Rooms: U165, U166, U167, U171, U172.