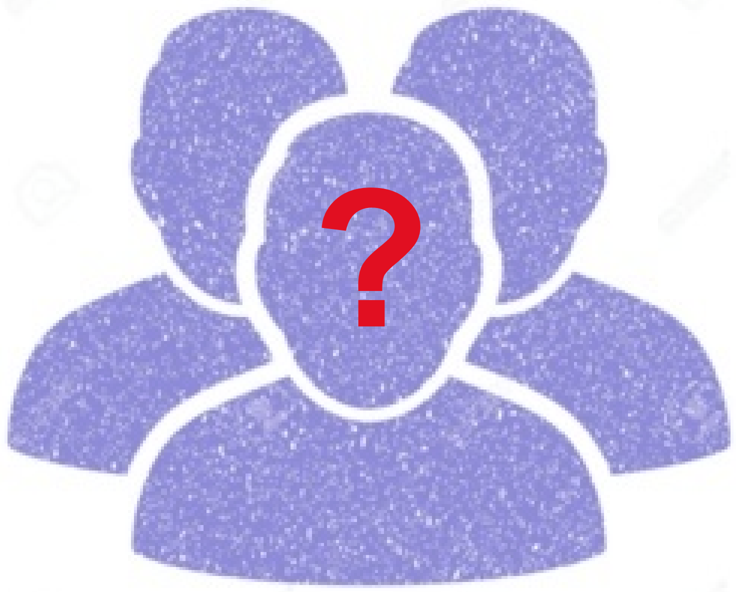
AI Booster – Week 02
Session 01 - Introduction

Franck JAOTOMBO & Sajad NAZARI
nazari@em-lyon.com

Welcome!

Pleased to meet you !

Few words about yourselves



Sajad Nazari

- Professor of AI and Applied Maths at EM LYON
- PhD in Computer science and in Mathematics
- https://em-lyon.com/en/sajad-nazari/ briefly

Interested in:

- **Data science (with Python)**
- **Knowledge representation**
- **ML**
- **Information science**

Outline & Program

- One week dedicated to improve your python programming skills and review basic statistical notions
- Day 1 (today!) => Introduction, data, data cleaning
- Day 2 => Univariate statistics
- Day 3 => Bivariate statistics
- Day 4 => Hypothesis testing and important distributions
- Day 5 => Review linear algebra

Teaching / learning materials

- Every day will follow the same schedule
- 1h30 of lecture (or less)
- 1h30 of in-class pratice (live coding session)
- Afternoon dedicated to practice (Tues., Wed., Thur. With a teaching assistant)
- Evaluation => individual quizz/exercices beginning of october + group project (at the end of week 3)

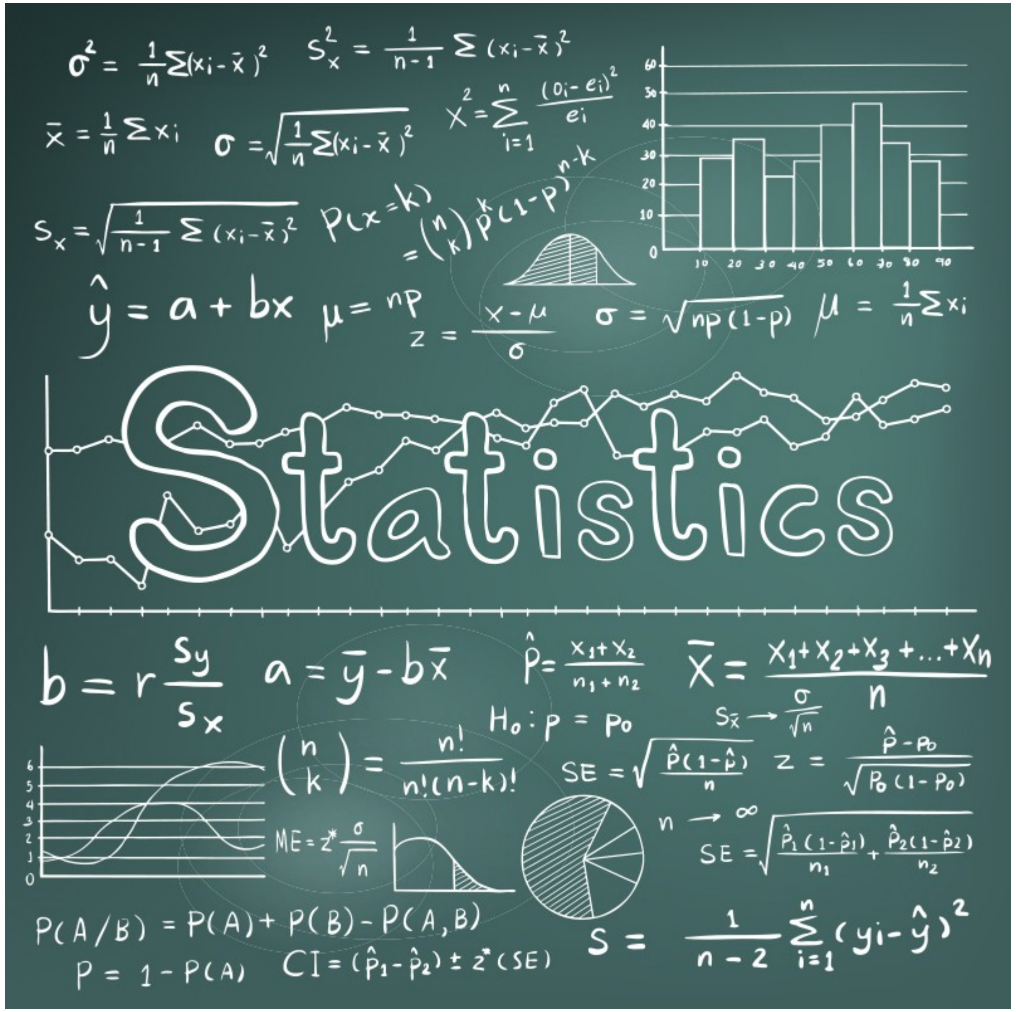
Why do we need statistics ?

General Introduction

Classification of variables

Making decision in an uncertain environment

- Running an organization (leading, managing, organizing…) is mostly about making decisions
  - Should we launch this new product on this market ?



Making decision in an uncertain environment

em lyon business school

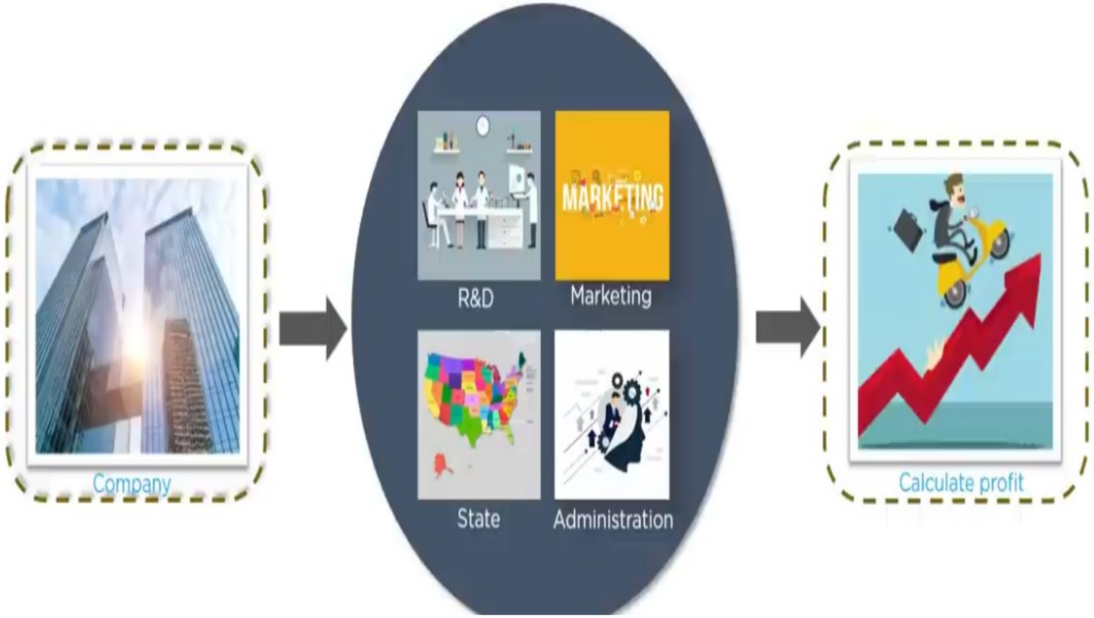
- **concrete example :**
- There are several of companies
- You want to invest in one of these companies
- **Key question :** how much money I could expect to earn ? Apart from the expenses
- **How ?**
  - By prediction : while you already know the expenses
- **Based on what ?**
  - former data from other companies containing expenses and profits

Making decision in an uncertain environment

Making decision in an uncertain environment

em lyon business school

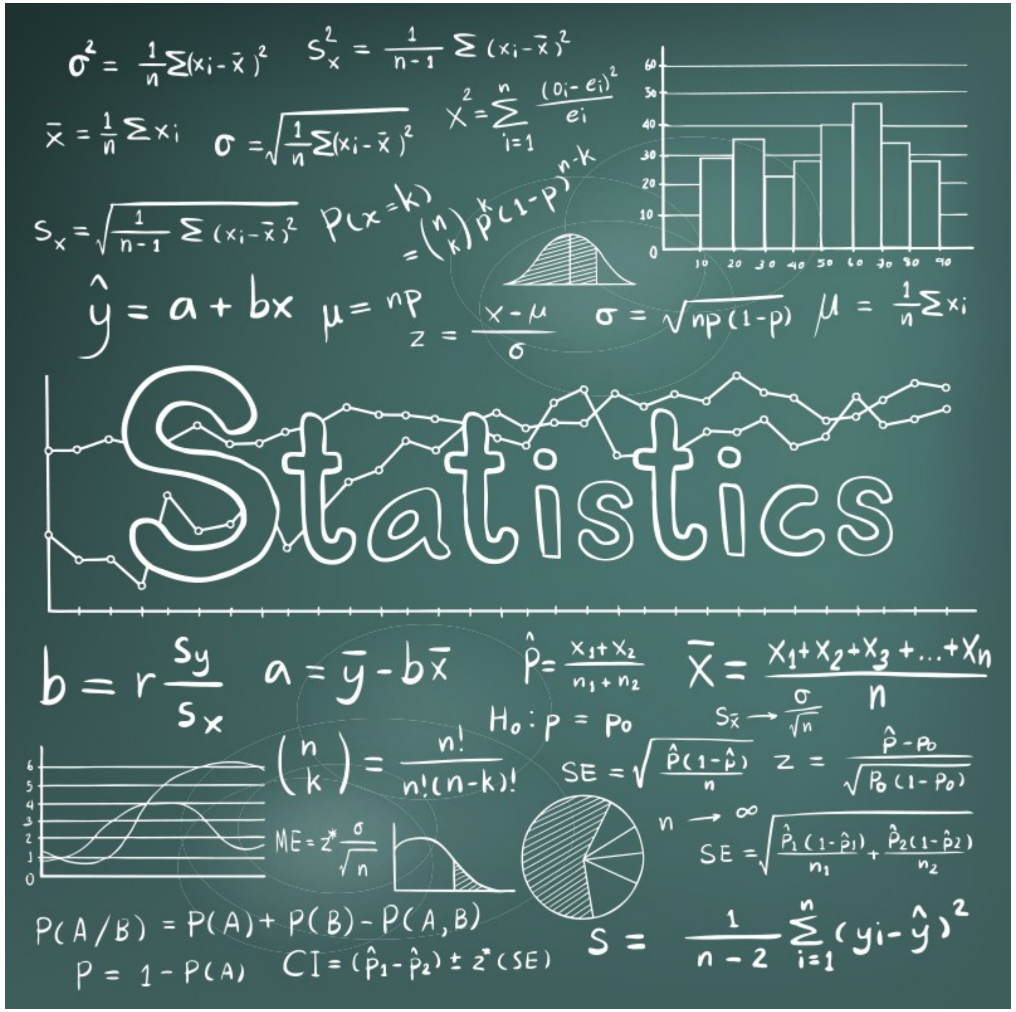
We have different expenses according to them we calculate the profit



	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

Making decision in an uncertain environment

- Running an organization (leading, managing, organizing…) is mostly about making decisions
  - Should we launch this new product on this market ?
- To make informed (wise) decisions, we need reliable information
  - Information is encapsulated within all sorts of data
  - Statistics is a tool to help process, summarize, analyze, and interpret data
- In sum : **statistics facilitates decision making**
- Another reason : this is the age of Artificial Intelligence
  - AI is largely based on statistics



Descriptive and Inferential Statistics

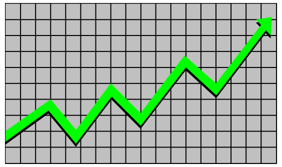
Two branches(applications) of statistics:

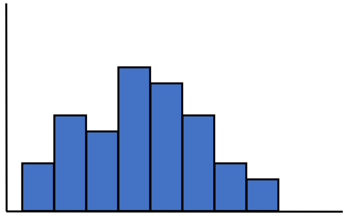
- Descriptive statistics
  - Graphical and numerical procedures to summarize and describe data
- Inferential statistics
  - Using data to make predictions, forecasts, and estimates to assist decision making
  - Project the information into a larger group

Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs





- Summarize data
  - e.g., Sample mean = $\frac{\sum X_i}{n}$

- Descriptive statistics
  - Graphical and numerical procedures to summarize and describe data

Inferential Statistics

em lyon business school

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 70 kgs
- Regression Analysis
  - e.g., Predicting house prices based on square footage

- Inferential statistics
  - Using data to make predictions, forecasts, and estimates to assist decision making
  - Project the information into a larger group

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Basic Vocabulary of Statistics

- EXAMPLE 1 : Men over 50 can lose weight on the pancake diet!
- EXAMPLE 2 : The academic performance of EM Lyon students, measured by their GPA

- POPLULATION
  - A population consists of all the items or individuals about which you want to draw a conclusion. The population is the "large group"
    - All men over 50 years old
    - All students enrolled at EM Lyon

- SAMPLE
  - A sample is the portion of a population selected for analysis. Time and money are limiting factors to collect all the data. The sample is the "small group"
    - 200 men over 50 years old participating in the survey
    - The students in this class

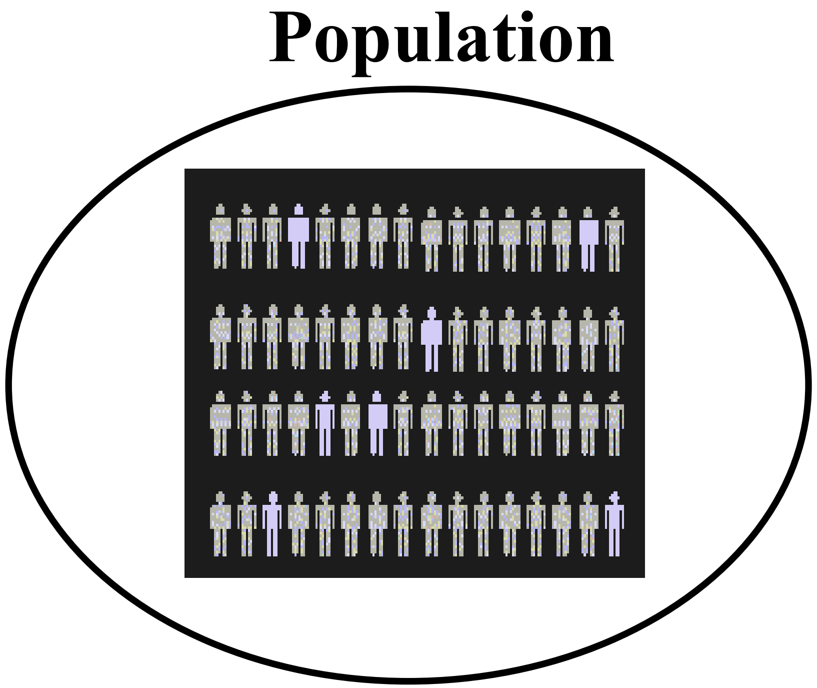
- Statistical (individual) unit or record
  - A single piece of data or a unit of the population
  - Usually a line (a row) in the data set
    - Mr. x
    - Mrs. y

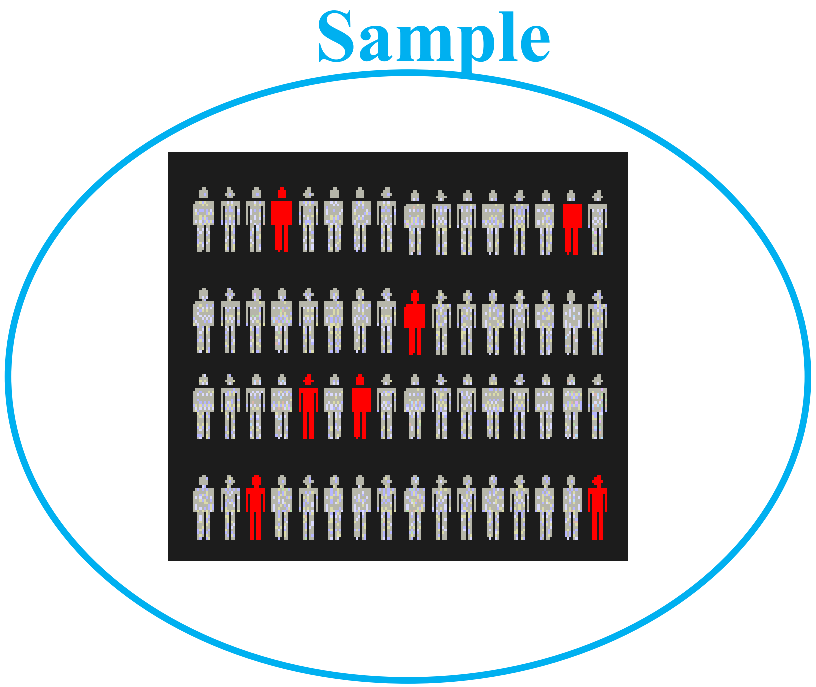
Basic Vocabulary of Statistics

- EXAMPLE 1 : Men over 50 can lose weight on the pancake diet!
- EXAMPLE 2 : The academic performance of EM Lyon students, measured by their GPA

- STATISTIC
  - A statistic is a numerical measure that describes a characteristic of a sample.
    - Average weight loss of the 200 men
    - Average GPA of the students of this class
- PARAMETER
  - A parameter is a numerical measure that describes a characteristic of a population.
    - **True average weight loss of all men over 50 years old**
    - **The variance of GPA of all students at EM Lyon**
- VARIABLES (Attribute in data sets)
  - Variables are characteristics of an item or individual that can vary from one unit to the next; they are what you analyze when you use a statistical method.
  - Usually a column in the data set
    - **Diet type, weight loss**
    - Major of study (e.g., Business Administration, Finance, Marketing) and GPA (Grade Point Average) of the student
- DATA
  - Data are the different values associated with a variable.
    - Recorded weights before and after the diet for each of the 200 men
    - Recorded GPAs and majors of the students of this class

Population vs. Sample

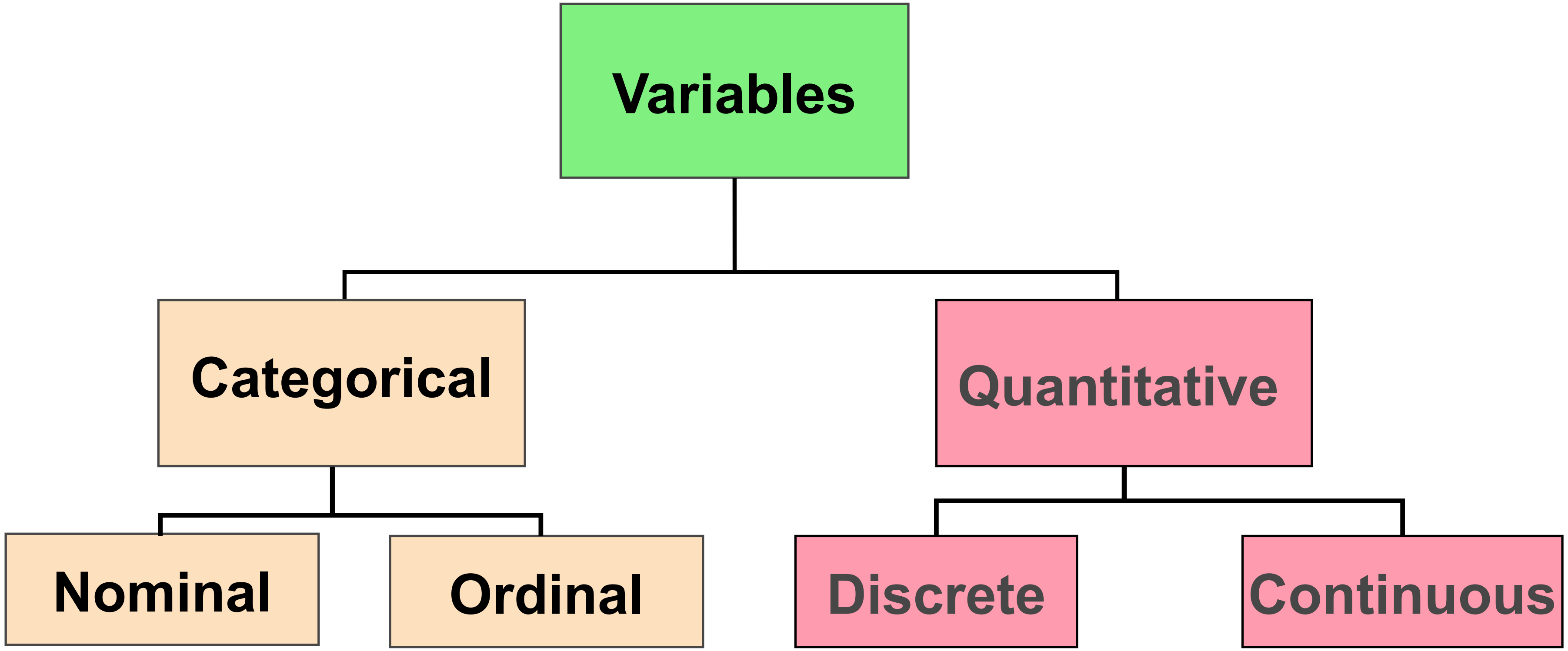




Measures used to describe the population are called **parameters**

Measures used to describe the sample are called **statistics**

Classification of Variables



Examples:

- Marital Status
- Blood Type
- Eye Color

(Defined Categories)

Examples: Ratings

- Good, Better, Best
- Low, Med, High

(Ordered Categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

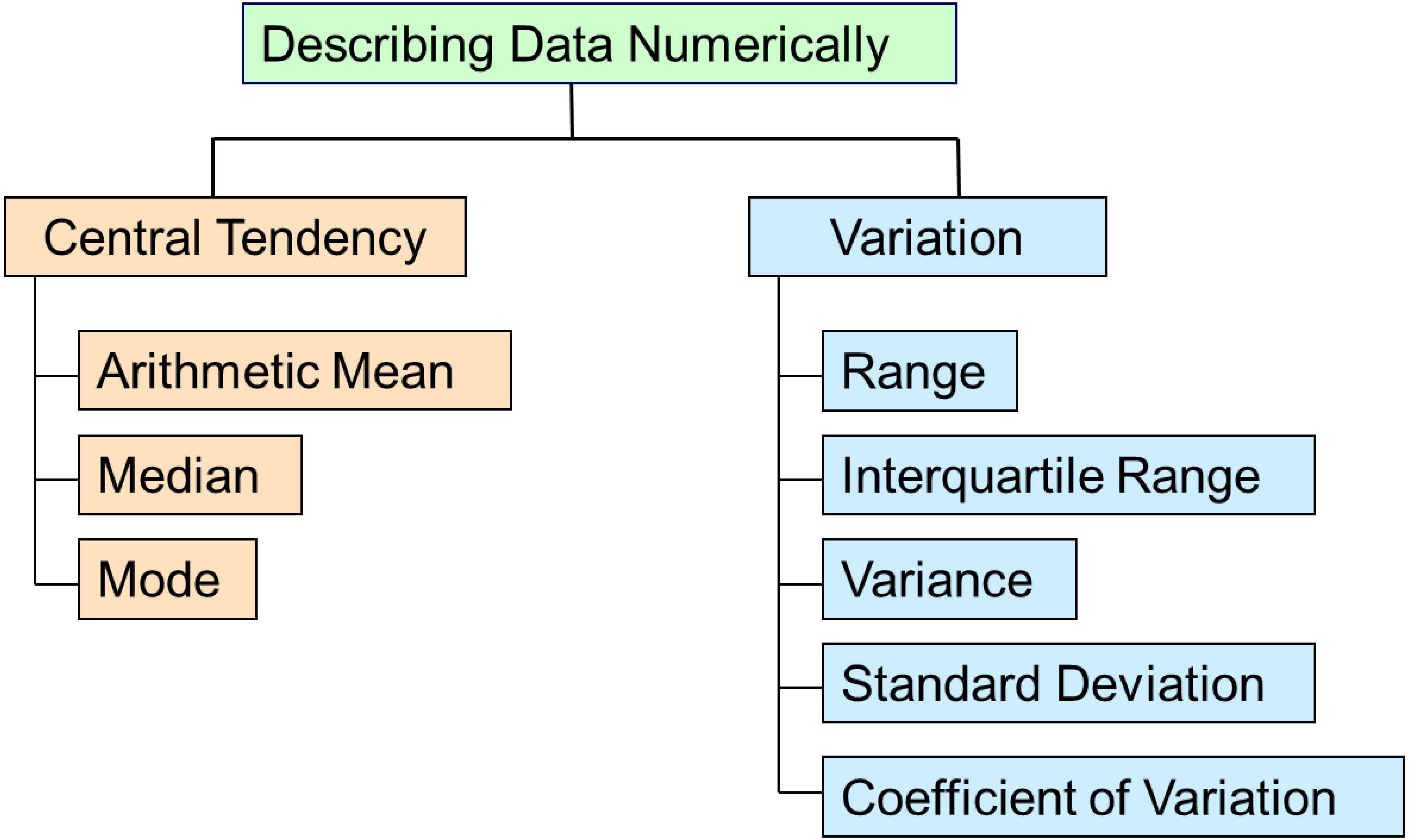
(Measured characteristics)

Describing Data Numerically

Central Tendency

Shape

Spread / Dispersion


Describing Data Numerically
Central Tendency
Arithmetic Mean
Median
Mode
Variation
Range
Interquartile Range
Variance
Standard Deviation
Coefficient of Variation

Measures of Central Tendency

- The **central tendency** is the extent to which all the data values group around a typical or central value. (salary)

- The **variation (spread / dispersion)** is the amount of dispersion or scattering of values

- The **shape** is the pattern of the distribution of values from the lowest value to the highest value.

Measures of Central Tendency

Central Tendency

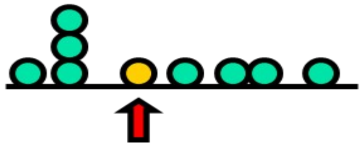
- Mean
- Median
- Mode

Mean

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

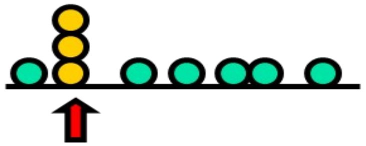
Arithmetic average

Median



Midpoint of ranked values

Mode



Most frequently observed value

(if one exists)

Arithmetic Mean

- The arithmetic mean (mean) is the most common measure of central tendency
  - For a population of N values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Population values

Population size

  - For a sample of size n:

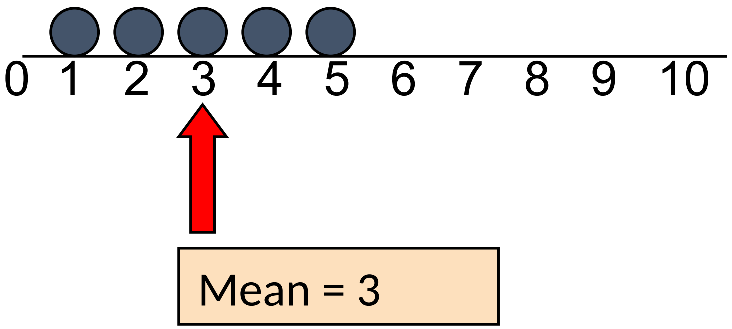
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Observed values

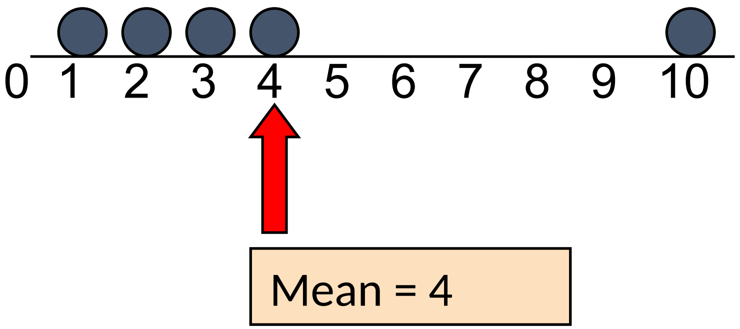
Sample size

Arithmetic Mean : example

- Mean = sum of values divided by the number of values
  - Affected by extreme values (outliers)
  - Solution : To get rid of outliers



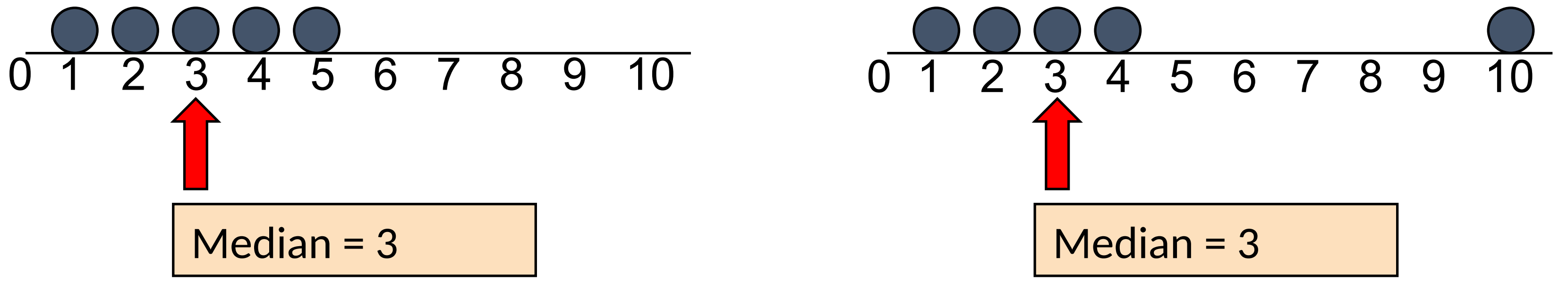
$$\frac{1+2+3+4+5}{5}=\frac{15}{5}=3$$



$$\frac{1+2+3+4+10}{5}=\frac{20}{5}=4$$

Median

- In an ordered list, the median is the "middle" number (50% above, 50% below)



  – Not affected by extreme values

Finding the Median

- The location of the median:

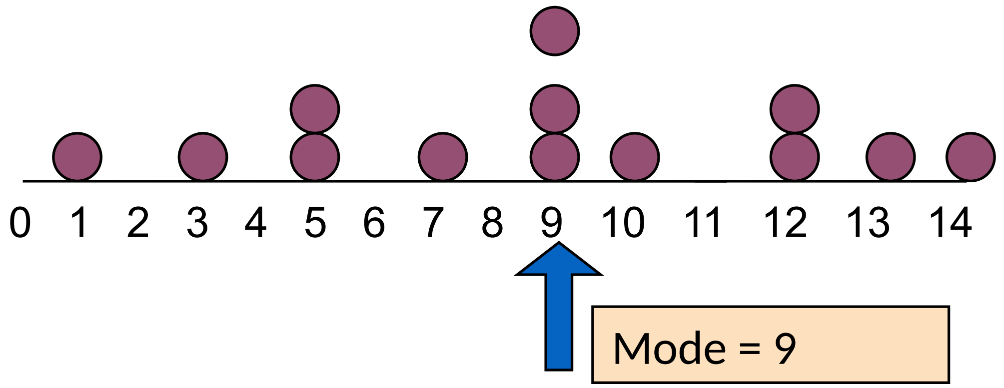
$$\text{Median position} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ position in the ordered data}$$

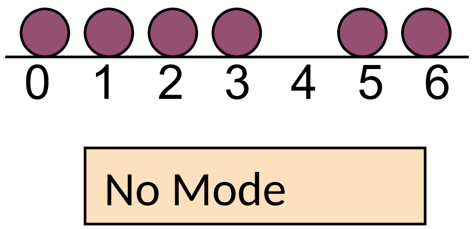
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

Mode

- A measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes: bimodal, trimodal, etc.





Example

em lyon business school

House Prices:	
	$2,000,000
	500,000
	300,000
	100,000
	100,000
Sum	3,000,000

- **Mean:** ($3,000,000/5)
  = **$600,000**

- **Median:** middle value of ranked data
  = **$300,000**

- **Mode:** most frequent value
  = **$100,000**

Geometric mean : example

An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two:

$$\underbrace{X_1 = \$100{,}000 \quad X_2}_{\text{50\% decrease}} = \$50{,}000 \quad \underbrace{}_{\text{100\% increase}} X_3 = \$100{,}000$$

The overall two-year return is zero, since it started and ended at the same level.

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

Arithmetic mean rate of return: $\overline{X} = \frac{(-.5)+(1)}{2} = .25 = 25\%$

Misleading result

Geometric Mean

- Geometric mean
  - Used to measure the rate of change of a variable over time

$$\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

- Geometric mean rate of return
  - Measures the status of an investment over time

$$\overline{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$

  - Where R_i is the rate of return in time period i

Geometric mean : example

An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two:

$$\underbrace{X_1 = \$100{,}000 \quad X_2}_{50\%\ \text{decrease}} = \underbrace{\$50{,}000 \quad X_3}_{100\%\ \text{increase}} = \$100{,}000$$

The overall two-year return is zero, since it started and ended at the same level.

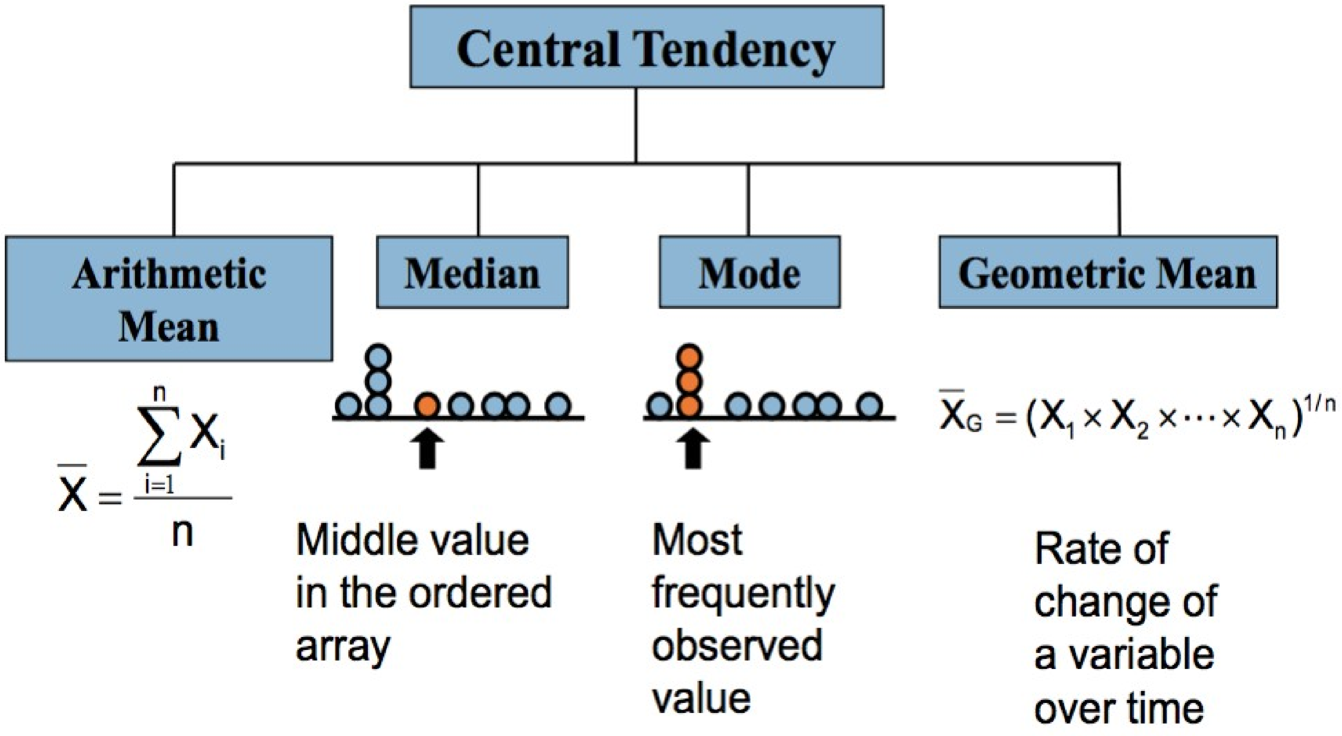
Geometric mean rate of return:

$$\begin{aligned}\bar{R}_G &= [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n}-1\\ &= [(1+(-.5))\times(1+(1))]^{1/2}-1\\ &= [(.50)\times(2)]^{1/2}-1 = 1^{1/2}-1 = 0\%\end{aligned}$$

More representative result

Summary : Central Tendency

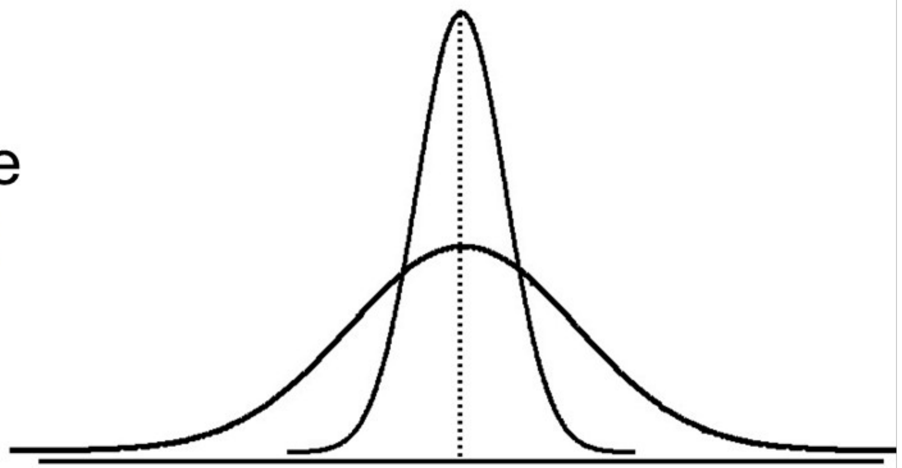
- The **mean** is generally used, unless extreme values (outliers) exist.
- The **median** is often used, since the median is not sensitive to extreme values. For example, median home prices may be reported for a region; it is less sensitive to outliers.
- In some situations it makes sense to report both the **mean** and the **median**.
- **Mode** in the only option for categorical data



Measures of Variability / Spread / Dispersion

- Variation
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

- Measures of variation give information on the spread or variability of the data values.



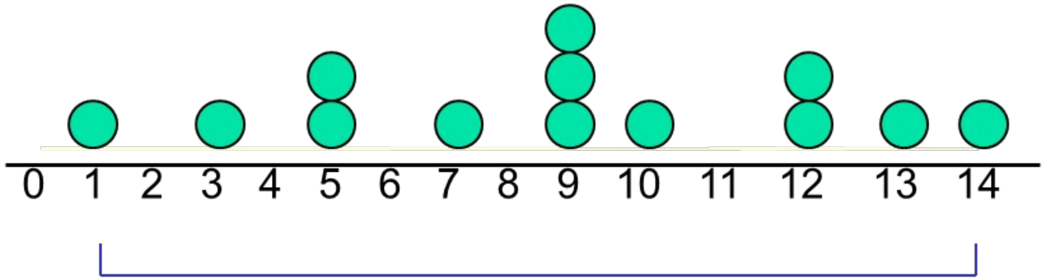
Same center, different variation

Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

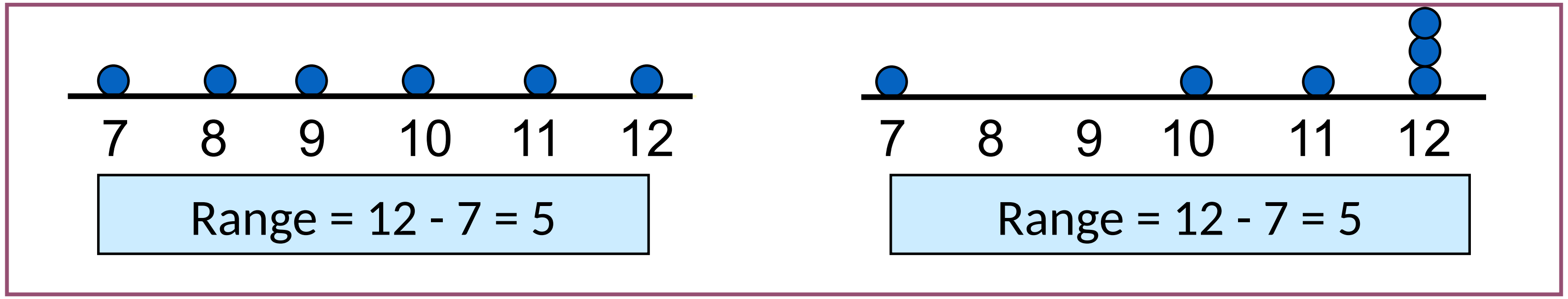
Example:



Range = 14 - 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed



- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

Interquartile Range

- Can eliminate some outlier problems by using the interquartile range

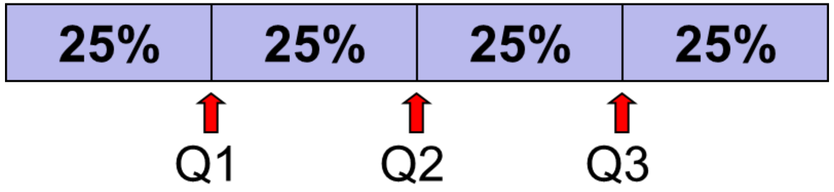
- Eliminate high- and low-valued observations and calculate the range of the middle 50% of the data

- Interquartile range = 3rd quartile – 1st quartile

$$IQR = Q_3 - Q_1$$

Quartile Measures

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger
- Q2 is the same as the median (50% of the observations are smaller and 50% are larger)
- The third quartile, Q3, is the value for which 75% of the observations are smaller and 25% are larger

- Find a quartile by determining the value in the appropriate position in the ranked data, where

  - First quartile position: $\mathbf{Q_1 = (n+1)/4}$ ranked value
  - Second quartile position: $\mathbf{Q_2 = (n+1)/2}$ ranked value
  - Third quartile position: $\mathbf{Q_3 = 3(n+1)/4}$ ranked value

where **n** is the number of observed values

Quartiles : example

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

n = 9

Q_1 is in the (9+1)/4 = 2.5 position of the ranked data,

so **Q_1 = position#2 + 0.5*(position#3 – position#2) = 12 + 0.5*(13-12) = 12.5**

Q_2 is in the (9+1)/2 = 5th position of the ranked data,

so **Q_2 = median = 16**

Q_3 is in the 3(9+1)/4 = 7.5 position of the ranked data,

so **Q_3 = position#7 + 0.5*(position#8 – position#7) = 18 + 0.5*(21-18) = 19.5**

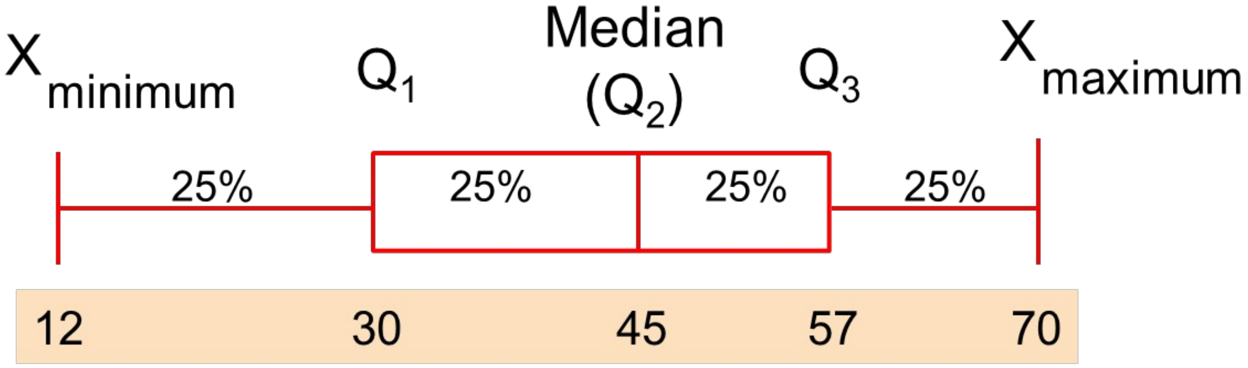
Q_1 and Q_3 are measures of non-central location
Q_2 = median, is a measure of central tendency

Introduction to Box-and-Whisker Plot

- A box-and-whisker plot is a graph that describes the shape of a distribution
- Created from **the five-number summary**: the minimum value, Q_1, the median, Q_3, and the maximum
- The inner box shows the range from Q_1 to Q_3, with a line drawn at the median
- Two "whiskers" extend from the box. One whisker is the line from Q_1 to the minimum, the other is the line from Q_3 to the maximum value

The plot can be oriented horizontally or vertically

Example:



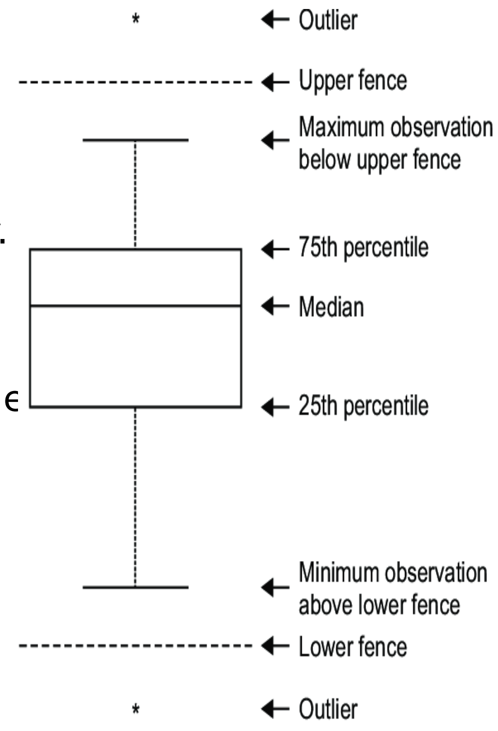
Constructing Full Boxplots

- Draw a single vertical (or horizontal) axis spanning the range of the data. Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box.

- Erect "fences" around the main part of the data.
  - The upper fence is 1.5 IQRs above the upper quartile.
  - The lower fence is 1.5 IQRs below the lower quartile.
  - Note: the fences only help with constructing the boxplot and should not appear in the final display.

Use the fences to grow "whiskers."
  - Draw lines from the ends of the box up and down to the most extreme data values found within the
  - If a data value falls outside one of the fences, we do not connect it with a whisker.

- Add the outliers by displaying any data values beyond the fences with special symbols.
  - We often use a different symbol for "far outliers" that are farther than 3IQRs from the quartiles.

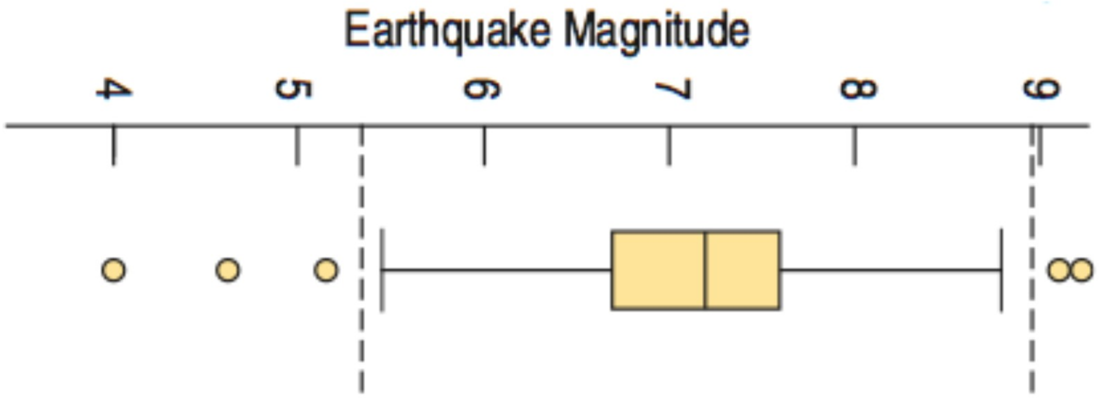


Boxplots : illustration

- The smallest tsunami-causing earthquake had magnitude 4.0 on the Richter scale.
- The largest tsunami-causing earthquake had magnitude 9.1.
- The middle half of tsunami-causing earthquakes is between 6.7 and 7.6.
- Half of tsunami-causing earthquakes have magnitudes below 7.2 and half are above 7.2.
- A tsunami-causing earthquake less than 6.7 is small.
- A tsunami-causing earthquake more than 7.6 is big.

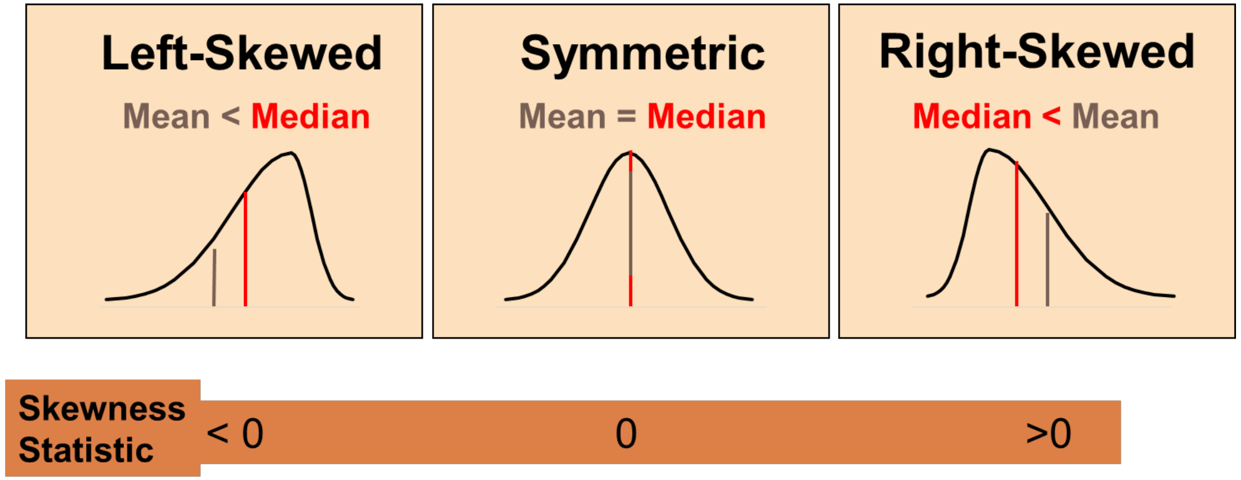
- Q1 = 6.7, Q3 = 7.6 so IQR = 7.6 – 6.7 = 0.9
- Lower Fence = 6.7 – 1.5 × 0.9 = 5.35
- Upper Fence = 7.6 + 1.5 × 0.9 = 8.95

Max	9.1
Q3	7.6
Median	7.2
Q1	6.7
Min	4.0



Shape of a distribution

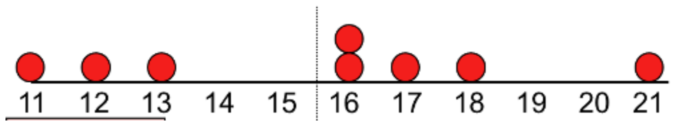
- Describes how data are distributed
  - elongated tail determines the direction of skew
  - The diagram is a continuous bar chart
- Two useful shape related statistics are:
  - Skewness
    - Measures the amount of asymmetry in a distribution



Variance

em lyon business school

Population Variance



Sample Variance

- Average of squared deviations of values from the mean
  - Population variance:

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

Where μ = population mean

N = population size

x_i = i^{th} value of the variable x

- Average (approximately) of squared deviations of values from the mean
  - Sample variance:

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

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

The bigger the variance is the more spread out the data is
The smaller the variance is the closer the data is

Standard Deviation

Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

  - Population standard deviation:

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

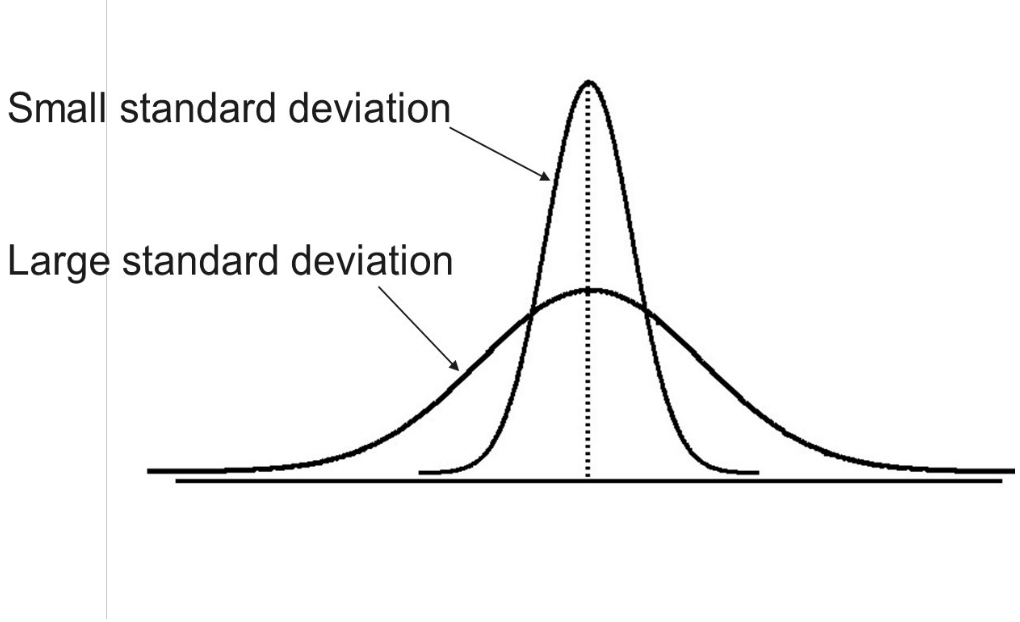
Sample Standard Deviation

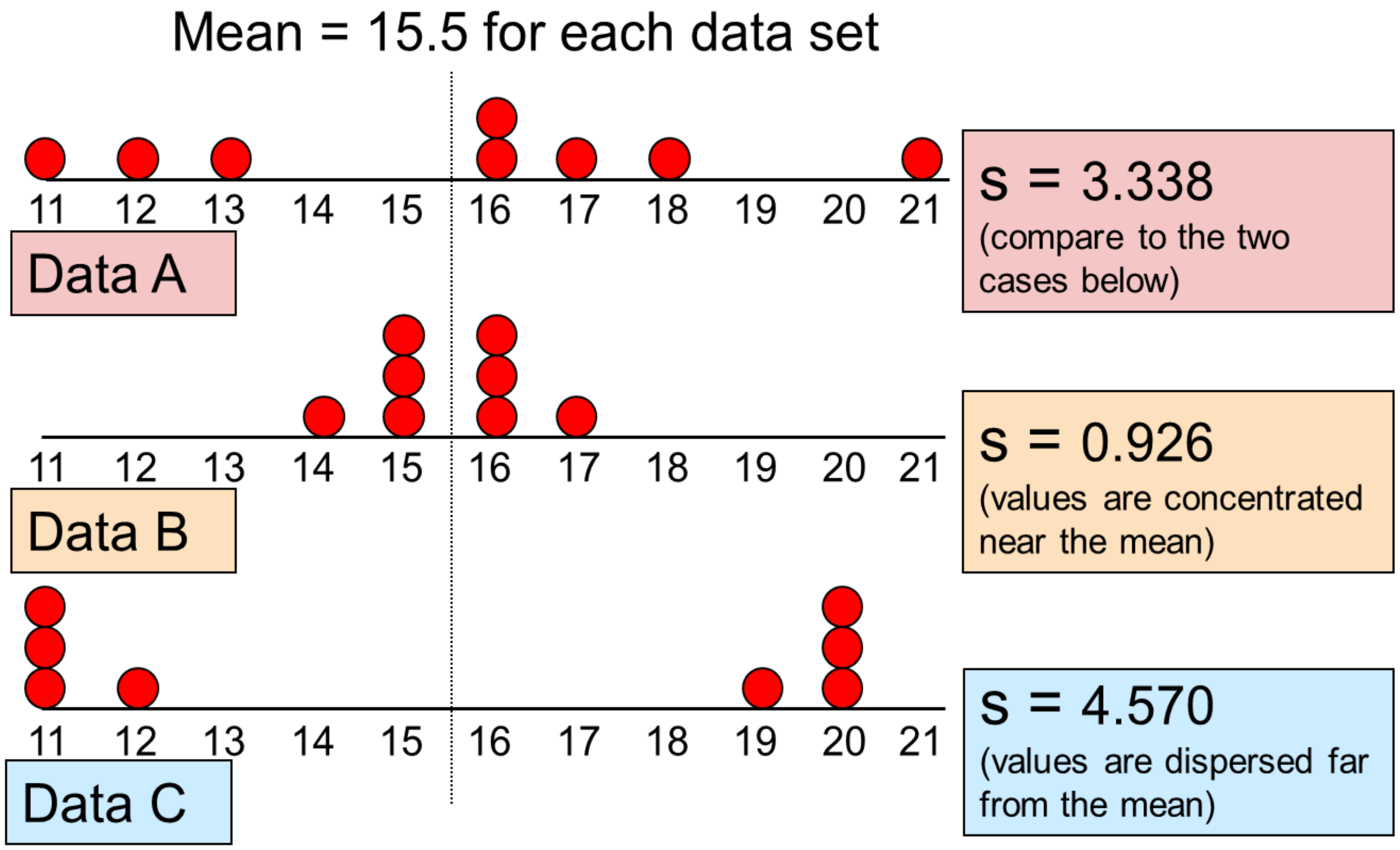
- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

  - Sample standard deviation:

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

Comparing standard deviation





Variability : summary

- The more the data are spread out, the greater the range, variance and standard deviation.
- The more the data are concentrated, the smaller the range, variance and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.

Comparing variation : Coefficient of Variation

- Question?
- Consider two cities A and B
  - The average house price and the standard deviation for a sample of both cities are respectively as follows:
    - $Mean_A = 1000000$ € and $SD_A = 10000$ €
    - $Mean_B = 12000$ € and $SD_B = 1000$ €
  - Which has more spread?

Comparing variation : Coefficient of Variation

em lyon business school

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right) \cdot 100\%$$

Sample coefficient of variation:

$$CV = \left(\frac{s}{\bar{x}}\right) \cdot 100\%$$

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Data cleaning

General Introduction

Classification of variables

Data cleaning

- **To generate meaningful, valid and reliable insights, clean, high-quality data is needed before analysis**
- We need to be able to reliably inform decision makers, the extent to which clean data reflects the reality of the data source - **Quality decisions rely on quality data**
- If data is incorrect, analysis and outcomes of analysis are unreliable even though they may appear correct
- Tools built using incorrect data e.g., algorithms, also become unreliable

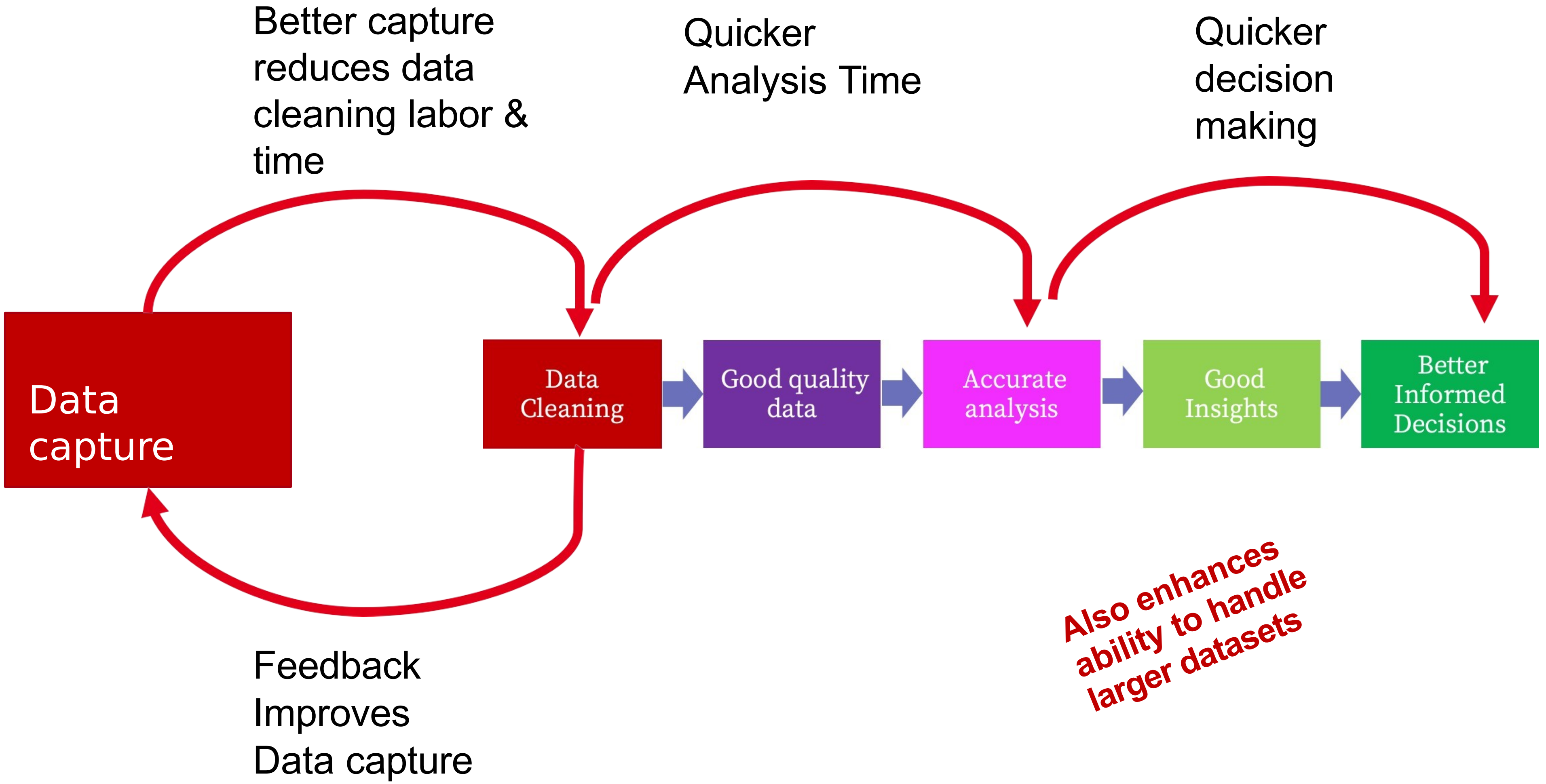
Data cleaning

em lyon business school

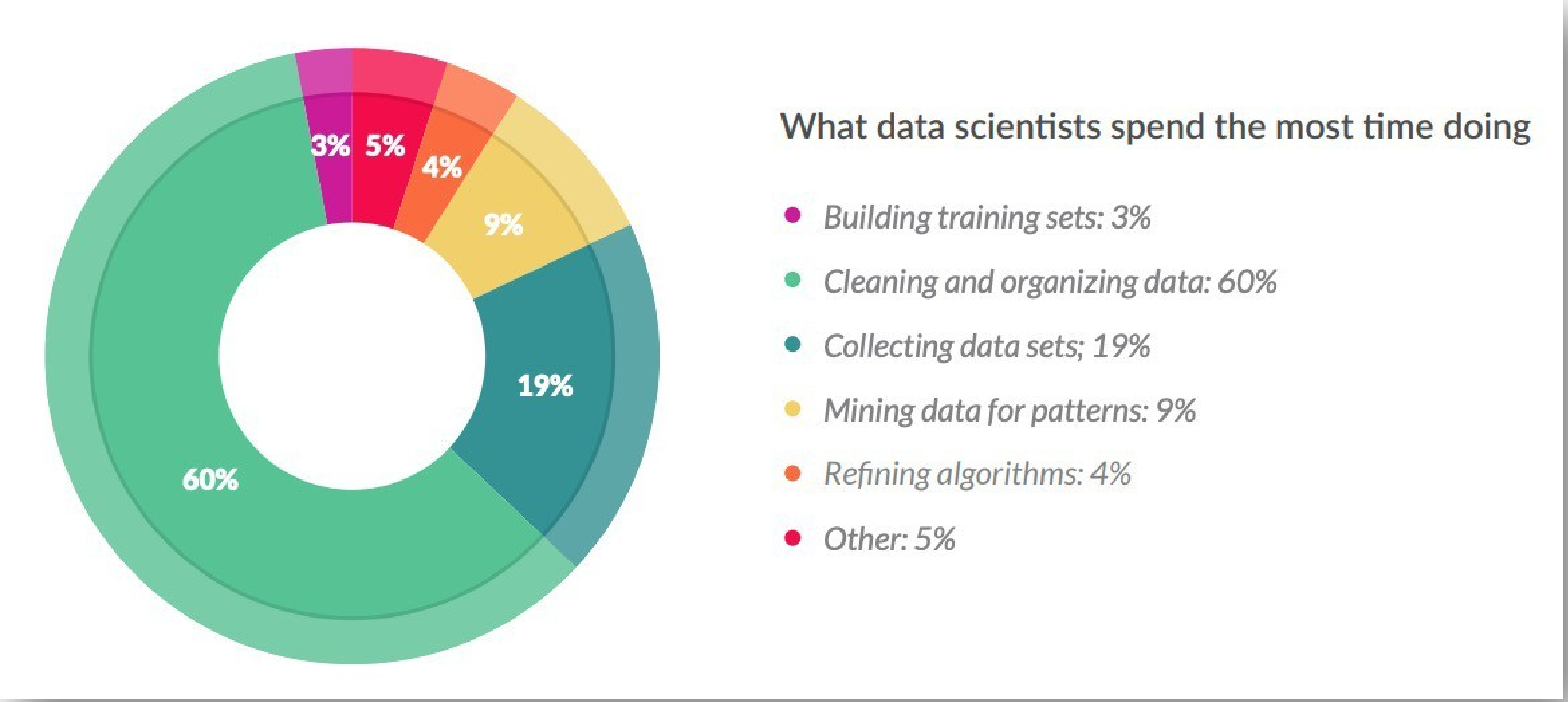
- Any changes between the original and clean data should be recorded and reported
- Steps taken in data cleaning can be used to provide feedback to improve data capture
- Helps to reduce delays during analysis

Data Cleaning → Good quality data → Accurate analysis → Good Insights → Better Informed Decisions

Data cleaning

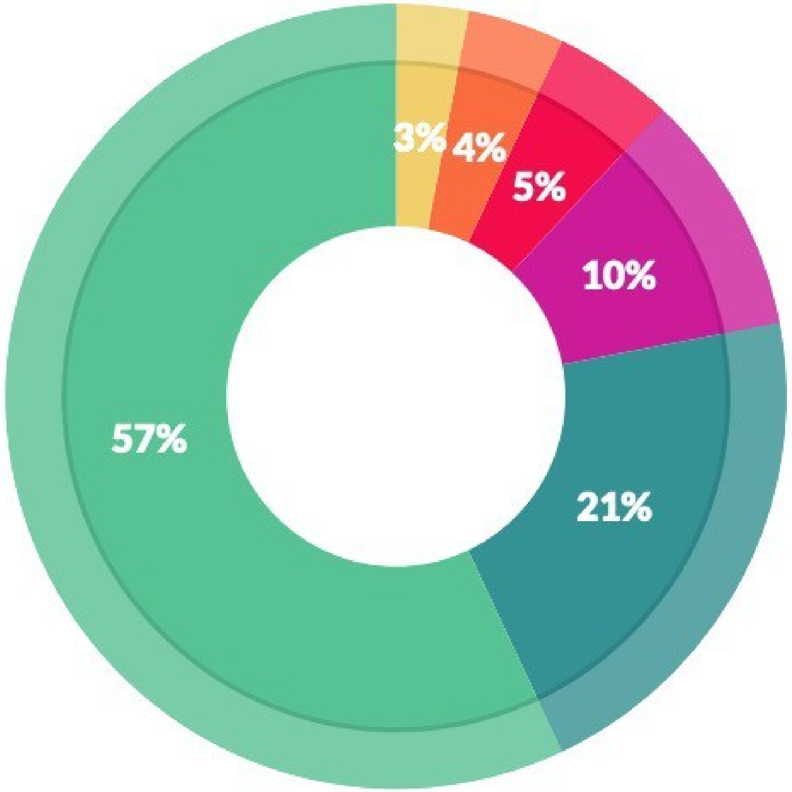


Data cleaning importance



https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Data cleaning is not attarctive !!



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

… this is the least enjoyable part of data science!

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Identifying irregularities in data

Broadly – steps taken in data cleaning consists of the following:

1. Backing up the original / master data
2. Data exploration: Understanding the data – its structure and dimensions
3. Detecting irregularities / errors in the data
4. Fixing the irregularities / errors in the data
5. Reporting your findings

Main kinds of irregularities

- Duplicated rows
- Missing values
- Outliers
- Spelling/typing mistakes
- Useless columns (duplicated or linearly dependant)
- Formating issues (. or , for fractionnal numbers, …)
- Multiple values in one column
- Encoding issues (utf vs latin1 vs iso)

Any questions ? +
Let’s start coding !