Statistics and Data Analysis

Responsible:
Dr. Tegawendé F. BISSYANDE
tegawende.bissyande@uni.lu

Course Author:

Médéric Hurier

Teacher Assistant:
Médéric Hurier
mederic.hurier@uni.lu

Course Features

- Sep. 20th - **Introduction to Big Data**

Part 1. Databases and Query Models for Big Data

- Sep. 27th - **Relational Databases: Reminders**
- Oct. 4th - **Relational Databases: Internals**
- Oct. 11th - **NoSQL & NewSQL Databases**
- Oct. 18th - **MapReduce Model**
- Oct. 25th - **Hadoop and Spark**
- Nov. 8th - **Datalog Model**

Part 2. Data Analysis and Machine Learning

- Nov. 15th - **Statistics and Data Analysis**
- Nov. 22th - **Communication and Visualization**
- Nov. 29th - **Features Engineering and Supervised Learning**
- Dec. 5st - **Unsupervised and Reinforcement Learning**
- Dec. 12th - **Homework Time**

Section Features

- Notions of statistics
- Datasets and dataframes
- Best practices on data management

Notion of statistics

Definition: Statistics

The science of drawing conclusions from data

1. derives knowledge from samples to population
2. establishes statistical significance of observed signal by studying randomness

How do scientist figure out whether something is good for you (e.g. video games, coffee)?

How do polls make accurate predictions based on data from only a small percent of voters ?

Descriptive and Statistical inference

- **Descriptive:** summarizing and describing data
  - goal: make description and comparison of datasets
- **Inference**: making conclusion from random samples
  - goal: generate inference and deduce relationships

Most common statistical measures

Centrality: mean, median, node

Dispersion: standard deviation (SD), IQR, range

Correlation: cross-tabulation and Pearson-r coefficient

Mean

English:

the sum of the values divided by the number values

Maths:

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$$\frac{2+4+5+9+10+0}{6} = \frac{30}{6} = 5$$

Note:

Zero values matters, don’t discount them !

Median

English:

The median is the midpoint of a distribution

Maths:

$$\Pr[R \leq x] \leq \frac{1}{2} \quad \text{and} \quad \Pr[R > x] < \frac{1}{2}.$$

Can be found by arranging the values from lowest to highest and picking the middle one

e.g. the median of [3, 3, 5, 9, 11] is 5

securityandtrust.lu

UNIVERSITÉ DU LUXEMBOURG

Mode

English:

The value in the set that occurs the most

= have the highest frequency

Example:

The mode of: [1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17] is 6

Note:

Can be used for both numerical and categorical data !

Standard Deviation (SD)

English:

Measures how far off the entries are from the mean

Maths:

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

Notes:

High SD value is often associated with high risk !

What about variance ?

The mean and SD have the same unit as the values

The variance is the square of the value unit

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

Example:

List: \$2, \$3, \$3, \$4, \$4, \$5, \$6, \$7, Mean=\$4.25

Variance=2.44 squared dollars, SD=\$1.56

Why is the SD useful ?

English:
No matter the list, the vast majority of entries

will be in the range average ± a few SDs

Maths: Chebychev's Inequality
A proportion of at least $1-1/k^2$ of the entries

will be in the range average ± $k \times$ SD

In any list:
1/9 of the entries are 3 or more SDs from the mean

Percentiles and Quartiles

The pth percentile of a list of numbers is the smallest number that is **at least** as large as the p% of the list

- 25th percentile: Lower/1st Quartile
- 50th percentile: Median (halfway point)
- 75th percentile: Upper/3rd Quartile

Example:
[0, 2, 4, 7, 9, 12]
<u>1st Quartile</u>= 2, <u>Median</u>= 7, <u>3rd Quartile</u>= 9

Interquartile Range (IQR)

English:
The difference between 3rd and 1st quartiles

Maths:

$$IQR = Q3 - Q1$$

Example:

1st Quartile=23, 3rd Quartile=31, IQR=31-23=8

Note:
Can be used to identify outliers (3 x IQR, John Tukey)

Distribution Range

English:

The difference between the largest and smallest values

Maths:

Range = maximum - minimum

Example:

<u>minimum</u>=15, <u>maximum</u>=45, <u>range</u>=30

Note:

most useful on small data sets (only two points)

Average of groups

It is not OK to take the average of averages !

The correct approach is to consider the group sizes

Example: a class has two sections

	average	section size	section proportions
Section 1	15	30	3/5
Section 2	8	20	2/5

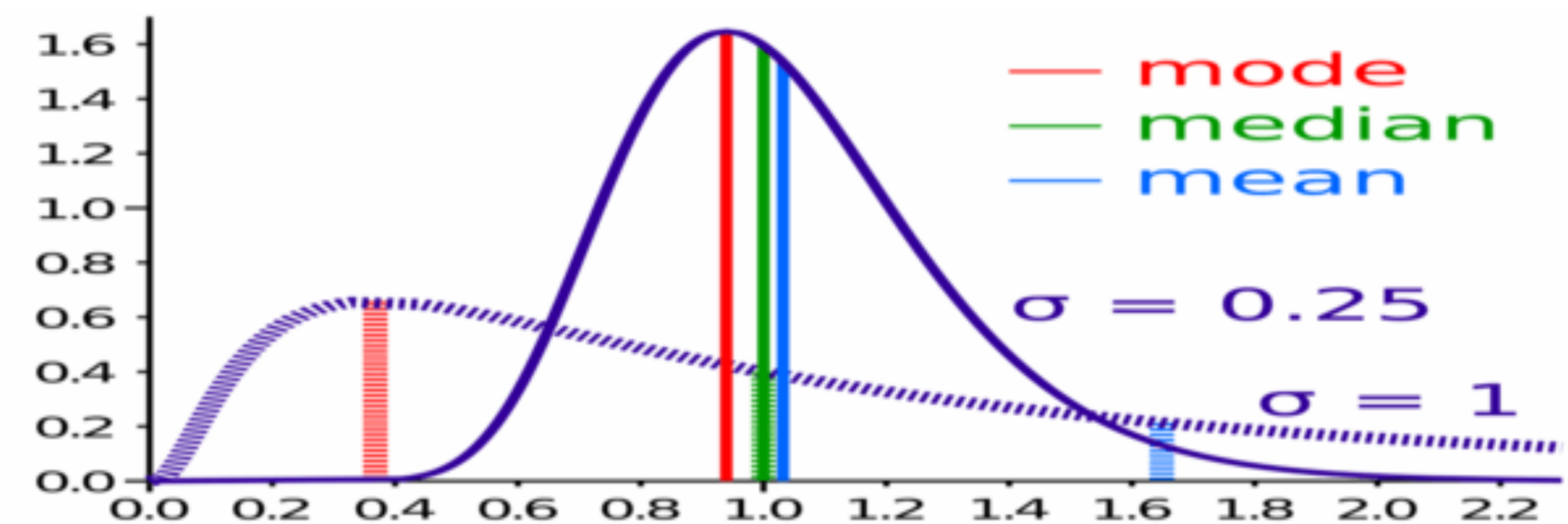
<u>Class average</u> = ((30 × 15) + (20 × 8)) / 50 = 12.2

<u>Average of averages</u> = (15+8) / 2 = 11.5

Robust Statistics

Metrics with good performance for data drawn from a wide range of probability distributions

e.g. not symmetric and with important outliers



The median and IQR are robust, not the mean and SD

What happens when Bill Gates enter a bar ?

Example: Student's test scores

If a student's test score is above average,

is the student in the top half of the class ?

Not necessarily

The class did well, but a few people did poorly

e.g. the mean is 65 and the median is 70.

Then a student who got 67 would be above average

but not in the top half of the class

Bottom-line

If you understand the concept well enough,

you don't need to do the calculation !

What is the mean and standard deviate of these list ?
[480, 480, 480, 500, 500, 500]

[0, 1]

This enable you to make quick estimations !

But try to avoid common traps !

age (years)	20-30	30-40	40-50	50-60	60-75	75+
average height(")	69.3	69.5	69.4	69.2	68.3	67.2

Intervals include the left endpoint but not the right.

[National Health and Nutrition Examination Survey, 1999-2002]

From this table:

Do men become shorter as they get older ?

NO !

This table is a snapshot of the population at a given time

Since these are not the same men, we cannot make conclusions

Cross Tabulation

A matrix format that displays the (multivariate) frequency distribution of the variables

They provide a basic picture of the interrelation between two variables and can help find interactions

	Right-handed	Left-handed	Total
Males	43	9	52
Females	44	4	48
Totals	87	13	100

Pearson-r coefficient

English:
A measure of the **linear** correlation between two variables

Math:

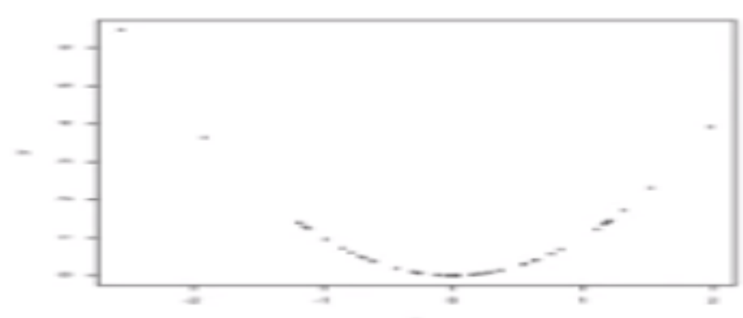
$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Notes:

- not robust (if outliers are present) !!!
- Correlation: 1 positive, 0 none, - 1 negative

Common mistakes

- **Correlation does not implies causation !**
  - e.g bigger the shoe size, better the kid can read
- **Pearson-s measures the linear association only !**



- **Correlated = Linearly related (only)**
  - in the plot above, variables may be related in another way
  - e.g. quadratic relation

Datasets and Dataframes

Definitions

A dataset is an actual files or collection of data

A dataframe is a memory represention of a dataset

More concretely, a **dataframe** is an index structure organized into named columns

Conceptually equivalent to:

- a table in a relational database
- an indexed matrix in mathematics

Types of datasets

- **Univariate**: contains only a single variable
  - interest: data distribution, shape, outliers ...
  - e.g. test scores of all students in all class
- **Bivariate**: a dataset containing two variables
  - interest: relationship between the variables
  - e.g. height and weight of students
- **Multivariate**: contains more than two variables
  - interest: find smaller group of variables to study
  - e.g. the form you filled at the beginning of the class

Continuous and Discrete variable

Continuous variables:
Values might be arbitrarily close to each other
in practice, can only measure up to a certain accuracy
e.g. height, weight, age

Discrete variables:
Values are separated from each other by fixed amounts
e.g. 0, 1, 2 ... are consecutive values separated by 1

It is possible to discretize or smooth variables

Main data types

Categorical: the values represent different categories
e.g. labels: fruits: apples, oranges …
do not have arithmetical meaning !

Ordinal: the values represent ordered categories
e.g. quality of meat: A, AA, AAA ...

Quantitative: the values represent numerical quantities
e.g. geoloc (interval, zero arbitrary), length (ratio, zero fixed)
do have an arithmetical meaning !

Are all numerical values quantitative ?

No !

Just because a variable has numerical values

doesn't mean it is quantitative

Example:

Computer ports, passenger class, rating ...

It doesn't make sense to do arithmetic on these numbers, they are just labels !

Operations associated to data type

Nominal: =, !=

Ordered: =, !=, <, >, <=, >=

Interval: =, !=, <, >, <=, >=, -
can measure distances or spans

Ratio: =, !=, <, >, <=, >=, -, /
can measure proportions (e.g. twice as much)

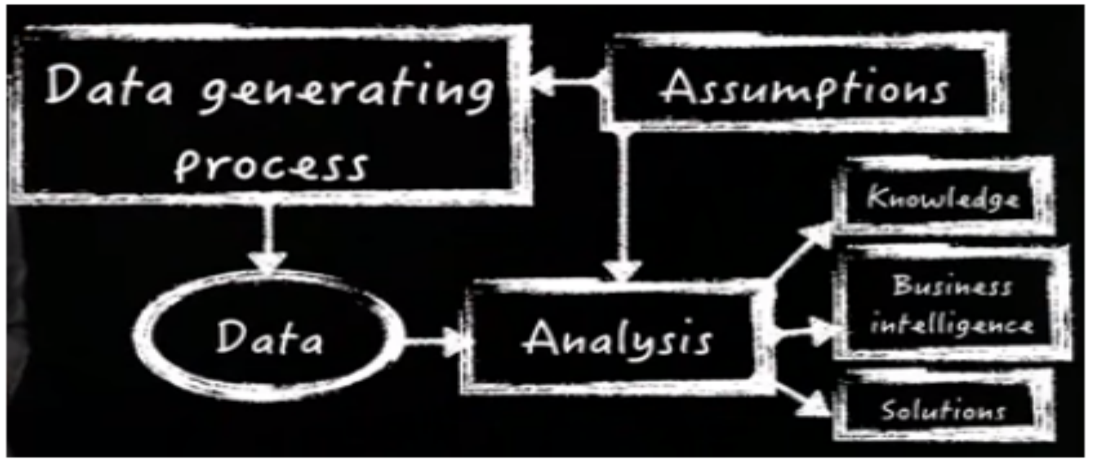
Best practices on data management

Data Science Process

Data is generated from a data generating process

this process may be under controlled or observed

The data scientist make assumptions on the process



feed the data to the analysis process to derive answers (knowledge, business intelligence, solutions)

Skills to be an efficient data analyst (1/2)

- **Learn a scripting language** (Perl, Python, Ruby): required for easy manipulation of data files and to eliminate overhead (boilerplate code)

- **Master regular expression**: required to deal with string and string like objects such as timestamps

- **Be comfortable browsing a database**: you should be able to use a command line/graphical frontend and figure out the schema/semantics easily

Skills to be an efficient data analyst (2/2)

- **Develop a good relation with your sysadmin/dba**: they can grant you access, create account, provide storage …, try to understand their position and constraints (they are paid to be paranoid !)
  - any production job has higher priority than an analysis !
- **Work on UNIX**: these systems were developed for precisely this kind of ad hoc programming with data
  - they continue to provide the most liberating environment for such work. They encourage you to devise solutions !
  - it develops your problem-solving attitudes !

Common sources for data in Enterprise

- **Databases**: contain data related to the business
  - OLTP (Online Transaction Processing = Production)
    - tend to be normalized, fast and busy
  - Data Warehouses (long term storage
    - tend to be denormalized and slow
- **Logfiles**: contains operational data (data activity)
  - usually contain much more information than databases
  - but deleted very quickly (e.g. less than two weeks)
- **Finance Department**: required for audit and tax
  - information is normative and therefore reliable

Advices to maintain a data collection

- **Make sure that all data sets are self-explanatory**
  - include metadata and all the information necessary
  - e.g. for time series, store the timestamp with the value
- **Make sure that all the analysis are reproducible**
  - keep track of the sources and transformations
- **keep data files readily available**: being able to run a script locally is better than waiting 12-24 hours
- **compress your data files** (e.g. gzip, tar.bz2 ...)
- **have a backup strategy**: get a second drive !

Recommendation for Data Format

- Use simple, portable and robust format
  - e.g. delimiter-separated text files, json files …
  - they can also be compress nicely !
- Keep metadata (either in the file or a directory)
  - be careful about additional payloads ! (e.g. XML files)
- Choose a format which is inexpensive to parse
  - again, XML files are notoriously expensive to parse

Know that the statistics communities use delimiter-separated text almost exclusively