COMP 333 — Week 2 Example 2: PISA

PISA

PISA is the OECD's Programme for International Student Assessment.

Every three years it tests 15-year-old students from all over the world in

- reading
- mathematics and
- science.

The latest results are for 2018.

The PISA web site `http://www.oecd.org/pisa/`

has a *Data* tab giving access to data from

2018, 2015, 2012, 2009, 2006, 2003, 2000

There is a document of "Insights and Interpretations" of the 2018 Results

`http://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf`

that shows a typical data analysis as would be done by "officials".

The graphs that the document shows are worth looking at.

PISA Example

Our Example 2 is presented in the video by Prof Patrick Meyer as supplementary material.

He looks at 2012 PISA data for the results in *mathematics*

and relates it to the PISA background data about access to computers in the household.

The presentation by Prof Meyer is a very good introduction to Descriptive Data Analysis.

He covers

- the types of data scales that arise as values for variables
- which statistical measurements make sense for these data scales
- the concept of outlier, and which statistics are *robust* to the presence of outliers in the data
- visualization of data (barchart, histogram, boxplot, scatter plot)
- univariate analysis, where you look at one variable
- multivariate analysis, that is, how to investigate relationships between two variables.

We will return to this topic under Descriptive Data Analysis in the future.

So the coverage here is more point-form than usual.

Data Scales

It is important to understand the type of data that has been collected for each variable.

This affects the descriptive statistics that you can use for the data

when you want to present central tendency or variation.

And it affects which plots make sense for the data!

It also affects how you can compare one variable to another.

Types of data scales are

- categorical data
  - nominal
  - ordinal
- continuous data
  - interval
  - ratio

You definitely need to learn which descriptive statistics and which plots make sense for each type of data scale!

Statistics

percentile

quartile

Central tendency: mode, median (P50, Q2), mean

Variation: range (=max-min), standard deviation, IQR (interquartile range) = Q3-Q1

Outliers and Extreme Values

Outliers are defined in terms of quartiles and IQR using 1.5*IQR
Extreme values are defined in terms of quartiles and IQR using 3.0*IQR
Boxplots show outliers and extreme values beyond the “whiskers”

Robust Statistics

Use median and IQR
Use boxplot

Visualizations

See examples of
barchart, and conditional barchart
histogram, and conditional histgram
boxplot, and side-by-side boxplot
scatter plot

Correlation

He mentions correlation
and how correlation as measured by Pearson coefficient is impacted by bivariate outliers

Next

You can see the details of this example of Descriptive Data Analysis in:

- the video by Professor Meyer

This is a very good video, well worth careful watching several times!

We will return to this topic under Descriptive Data Analysis in the future.