COMP 333 — Week 3 `scipy` Descriptive Statistics

It is very important to understand your data.

The field of *descriptive statistics* has developed an approach that describes

centre the "middle" point of the observed data

spread the variability on the data

shape the symmetry, skewness, number of peaks, and length of tails in the data

unusual features such as outliers and extreme values

You want to describe the data both quantitatively (with *statistics*) and visually (with *plots*).

READ the Article

The following article is provided as supplementary material.

It is a must READ

as it is an excellent coverage of descriptive statistics in Python, for `scipy`, for `pandas`, and for visualization.

Python Statistics Fundamentals: How to Describe Your Data by Mirko Stojiljkovic `https://realpython.com/python-statistics/`

Revise Statistics

You should use the opportunity of reading the article to revise your knowledge of statistics.
You need to *understand* the following terminology.
You do *not* need to remember formulas as you have access to the cheat sheets.

Terminology
observation, variable
population, sample
central tendency: mode, median, mean
variability: range, variance standard deviation, IQR (interquartile range)
percentile, quartile

And to pick up new knowledge:

outlier, extreme value
skew
kurtosis
correlation, covariance, Pearson correlation coefficient

Robust Statistics

You want the description to be *robust* in the sense that
their values are not greatly affected by the presence or absence of outliers.

The robust statistics are defined in terms of the *quartiles* Q_1, Q_2, and Q_3
with the *median* Q_2 used to describe the **central tendency**
and the *interquartile range (IQR)*, which is $Q_3 - Q_1$, for the **variability**.

Five Number Summary

Statistics has adopted the following five numbers as a robust quantitative description of a dataset:

- minimum,
- first quartile,
- median (the second quartile),
- third quartile, and
- maximum.

A traditional boxplot would use these five numbers to construct

the box (from Q1 to Q3) showing a line for the median Q2, and

the whiskers which end at the maximum and minimum value.

Today, a boxplot will indicate *fences* at the end of the whiskers for the start of outliers,

$Q1 - 1.5 \times IQR$

and

$Q3 + 1.5 \times IQR$

There may be a differentiation between outliers and *extreme values*, which start at

$Q1 - 3.0 \times IQR$

and

$Q3 + 3.0 \times IQR$

The box plot will show outliers and extreme values as small circles (dots)

and differentiate by having open circles and closed circles.

Python `scipy` `describe()`

The `scipy` describe() returns an object that holds the following descriptive statistics:

- **nobs**: the number of observations or elements in your dataset
- **minmax**: the tuple with the minimum and maximum values of your dataset
- **mean**: the mean of your dataset
- **variance**: the variance of your dataset
- **skewness**: the skewness of your dataset
- **kurtosis**: the kurtosis of your dataset

Kurtosis is a measure of *"tailedness"*
as explained in the wikipedia article `https://en.wikipedia.org/wiki/Kurtosis`
cited by Mirko Stojiljkovic.

Note that this does **not** provide robust statistics, but instead uses the mean and the variance.

Python `pandas` `describe()`

The `pandas` describe() returns a Series object that holds the following descriptive statistics:

- **count**: the number of elements in your dataset
- **mean**: the mean of your dataset
- **std**: the standard deviation of your dataset
- **min and max**: the minimum and maximum values of your dataset
- **25%, 50%, and 75%**: the quartiles of your dataset

Note that this provides **both** the robust and non-robust statistics.

Bivariate Descriptive Statistics

Your dataset will consist of many observations
and each observation will consist of values for many variables.

The *univariate descriptive statistics* above describe each variable individually.

The *bivariate descriptive statistics* describe the relationship between a pair of variables.

This is done in terms of

- *correlation*, whether there is a positive, negative, or weak relationship between corresponding pairs of values;
- *covariance*, the strength and direction of a relationship between a pair of variables;
- *correlation coefficient*, a measure of correlation.

Pearson Correlation Coefficient

The *Pearson product-moment correlation coefficient*,
often simply called the *correlation coefficient*,
is denoted by the symbol r.

It is measure of the correlation between the two variables.

It ranges from -1 for maximum negative correlation to +1 for maximum positive correlation.

A value of zero, or near zero, indicates weak correlation.

Bivariate Plotting

Visualization of the relationship between variables is useful.

The most common plot used for two variables is the *scatter plot*.
You can also use *side-by-side boxplots*.

When one of the variables is categorical, you can use *conditional* variations
of barcharts, histograms, and boxplots.
See Example 2 for some examples of these conditional plots.

Many bivariate plots can be organized into a *grid* to illustrate many variables at once.

What Else to Describe?

Looking ahead to Data Wrangling, Modeling, and Machine Learning, there may be things beyond the five number summary of robust statistics that are useful descriptions of the data:

- number of observations
- number of entries
- number of unique entries
- number of missing entries
- number of outliers
- number of extreme values