NPA Data Science

6: Data analysis and statistics

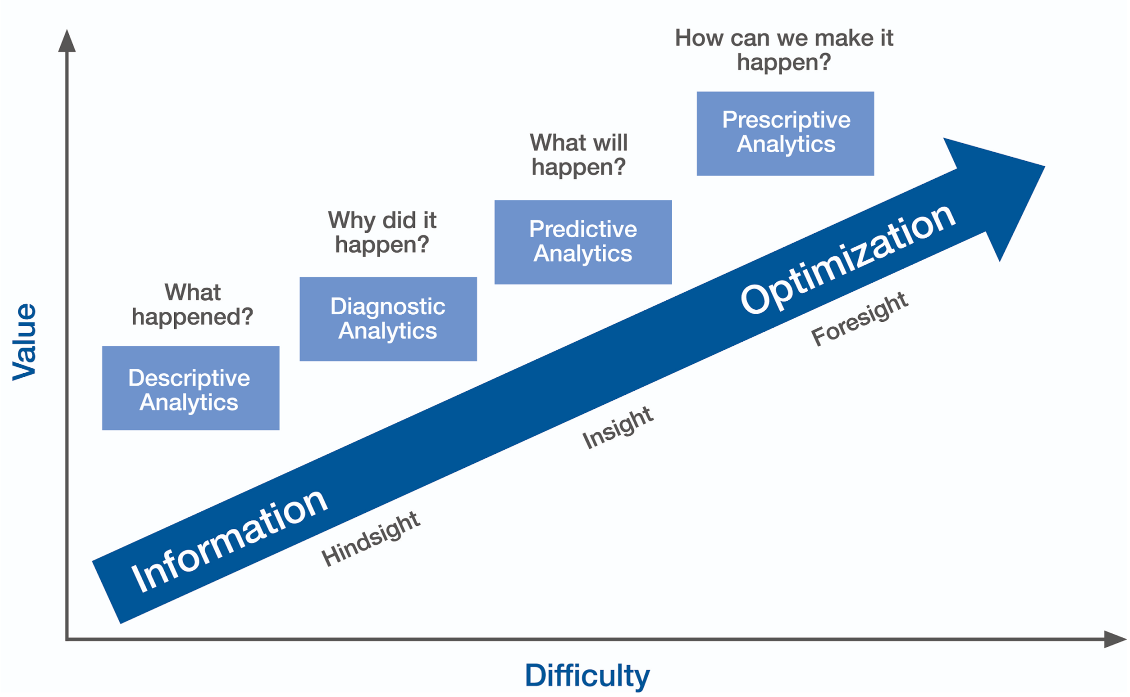
Topic Outcomes

- Explain how data can be analysed and the tools that can be used to perform analysis
- Explain descriptive analytics and predictive analytics
- Explain statistical techniques involved in data science

Data Analysis & Analytics

Data Analytics focuses on the entire methodology (i.e. the tools and techniques) for obtaining useful insights from data.

The four main types of analytics are descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics. These occur on a scale of difficulty and build on each other with more value to be gained from the data undergoing analysis as the difficulty level increases. The figure below shows that different kinds of analytics can produce knowledge of hindsight, insight and foresight, which is valuable for different organisations in many ways.



The two main types to be aware of are descriptive and predictive. Descriptive analytics tell you what is happening and predictive analytics tell you what is likely to happen.
Descriptive analytics are generally used to generate regular reports(daily, weekly, yearly, etc.) on specific parameters of the dataset and to display those reports through interactive dashboards or score-cards.
Predictive analytics, on the other hand, can take a dataset and make predictions based on past events or modeling.

To take the example of healthcare, with descriptive analytics, one could understand readmission rate, mortality rate, or average wait time at the pharmacy. With predictive analytics one may predict whether a patient is at a high risk of heart attack, which patient is likely to be read-mitted after a surgery. For this reason, predictive analytics in healthcare settings has received a great amount of interest. The knowledge gained through applying predictive analytics in health and medicine will change the way medicine is practiced while enhancing our ability to prevent and treat significant diseases and illnesses.

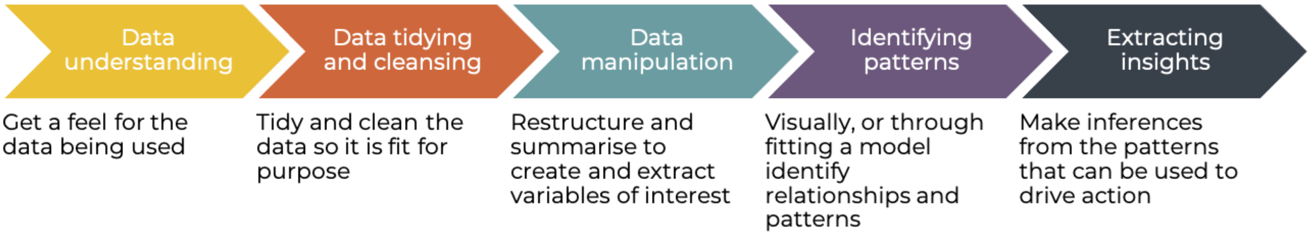
Analysing data

Data analysis is a subset of the data analytics methodology focused on compiling and reviewing data to aid in decision making. It is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making. This is done using analytical or statistical tools. Some of these tools are programming languages like R or Python. Microsoft Excel is also popular in the world of data analytics. However, Microsoft Excel isn't meant for data analysis, but it can still handle statistics. It was developed for Microsoft by Apple as an accounting package for Microsoft’s new windows operating system in 1987.

There are two main categories for different types of data: Quantitative and Qualitative. Quantitative refers to numerical and metric data, i.e. numbers and that which can be quantified. Qualitative data refers to textual data i.e. words. Types of analysis are usually dictated by the type of data that you are working with. Examples of types of data include Numerical, categorical, textual, Boolean, geographic, demographic, pictorial, visual, audio.

It is generally accepted that around 80% of any data analyst’s time is spent cleaning and manipulating data. These activities are both important and time consuming. The other activities involve detailed understanding of the dataset before any manipulation and ensuring that conclusions are drawn, and actions are taken at the end of the process.

There is a structured approach to carrying out a data analysis, which, if followed, will minimise mistakes and maximise the validity of the conclusions or insights extracted from the data. This is what would be done within PPDAC’s analysis step:


Data understanding
Get a feel for the data being used
Data tidying and cleansing
Tidy and clean the data so it is fit for purpose
Data manipulation
Restructure and summarise to create and extract variables of interest
Identifying patterns
Visually, or through fitting a model identify relationships and patterns
Extracting insights
Make inferences from the patterns that can be used to drive action

Simple Strategies for Analysis

Averages (mean, median and mode), sampling and surveying, and data errors (false positives and false negatives).

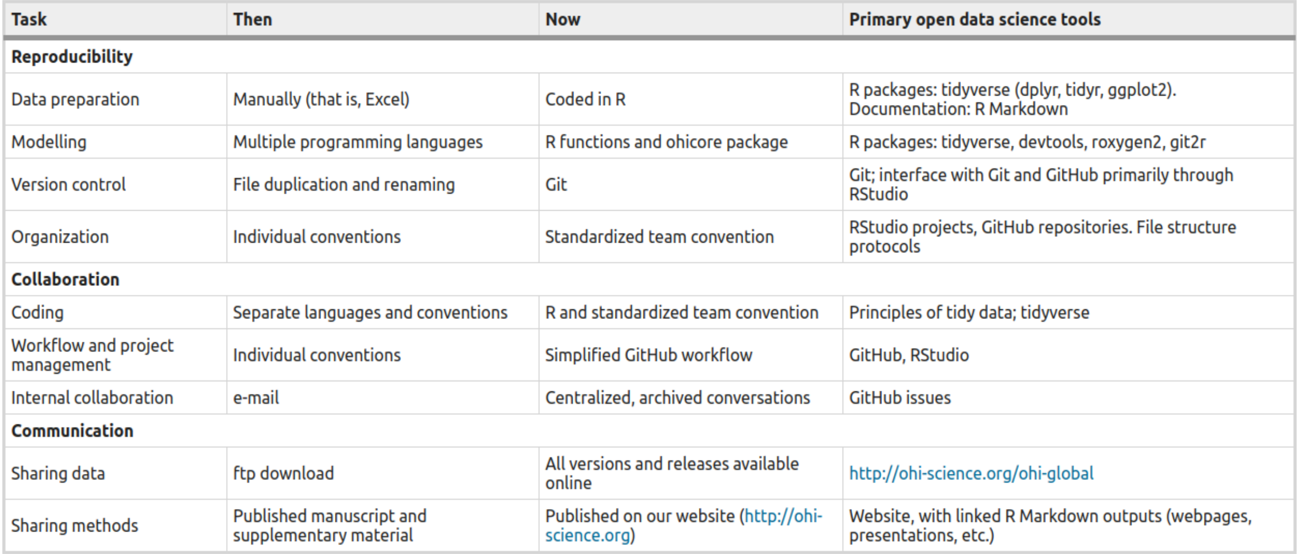
Mean	The arithmetic mean is the sum of all the values in the dataset, divided by the number of dataset points. Care should be taken when missing values are present. The mean value is very sensitive to extreme or outlying values.
Median	The median is the middle value of an ordered dataset. If there is an odd number of points the middle value is used. If there is an even number the mean of the two middle values is used. The median is not affected by outliers.

Mode	The mode is the most frequent value, this can be used with both numeric and non-numeric sata. For continuous frequency distributions it is the maximum value. There can be more than one mode if the distribution has multiple peaks.

Tools for Data Analysis

With the improvements in the existing tools and entry of newer ones into the Data Science scene, many tasks have become achievable, which were earlier either too intricate or unmanageable. The core idea behind these tools is to unite data analysis, machine learning, statistics and related concepts to make the most out of data. These tools are critical for anyone looking to dive into the world of Data Science and picking the right tools can make a world of difference. While some tools are worthy of being called all-rounders, some cater to specific niches.

Whilst there are many proprietary (commercial) tools for data analysis, there increasingly exist powerful tools which are freely available to use. This is important for the advancement of open data science, but may often involve a transition process to using 'open data science tools'. For example, Lowndes et al. (2017) explain the incremental process of adopting new tools necessary for their work as environmental scientists, shown in the image below. They explain how this led to better science based on transparency, openness and reproducibility and was more efficient and productive.



Task	Then	Now	Primary open data science tools
Reproducibility			
Data preparation	Manually (that is, Excel)	Coded in R	R packages: tidyverse (dplyr, tidyr, ggplot2). Documentation: R Markdown
Modelling	Multiple programming languages	R functions and ohicore package	R packages: tidyverse, devtools, roxygen2, git2r
Version control	File duplication and renaming	Git	Git; interface with Git and GitHub primarily through RStudio
Organization	Individual conventions	Standardized team convention	RStudio projects, GitHub repositories. File structure protocols
Collaboration			
Coding	Separate languages and conventions	R and standardized team convention	Principles of tidy data; tidyverse
Workflow and project management	Individual conventions	Simplified GitHub workflow	GitHub, RStudio
Internal collaboration	e-mail	Centralized, archived conversations	GitHub issues
Communication			
Sharing data	ftp download	All versions and releases available online	http://ohi-science.org/ohi-global
Sharing methods	Published manuscript and supplementary material	Published on our website (http://ohi-science.org)	Website, with linked R Markdown outputs (webpages, presentations, etc.)

Examples of some open data science tools for data analysis include:

RStudio https://rstudio.com/

RStudio is dedicated to sustainable investment in free and open-source software for data science, to help people understand and improve the world through data.

ApacheSpark https://spark.apache.org/

Apache Spark by Apache Software Foundation is a tool for analyzing and working on large-scale data. It allows you to program clusters of data for processing them collectively by incorporating data parallelism and fault-tolerance. For data clusters, Spark requires a cluster manager and a distributed storage system. Spark also inherits some of the features from Hadoop, such as YARN, MapR and HDFS.

Hadoop https://hadoop.apache.org/

Hadoop is an open-source software by Apache Software Foundation authorised under the Apache License 2.0. By using parallel processing across clusters of nodes, it facilitates solving complex computational problems and data-intensive tasks. Hadoop does this by splitting large files into chunks and sending it over to nodes with instructions.
KNIME: https://www.knime.com/
KNIME is a multi-purpose tool that does data reporting and analytics while enabling easy integration of elements such as data mining and machine learning onto your data. KNIME's intuitive GUI allows for easy extraction, transformation and loading of data with minimal programming knowledge. Enabling the creation of visual data pipelines to create models and interactive views, KNIME can work on large data volumes.

Statistics for Data Science

William Oxbury, GCHQ said in 2018, "the business of analysing and utilising data is fundamentally the business of statistics".
Data science is often considered to be applied statistics, as it involves the application of statistics to real-world problems. However, data science also involves implementing a solution or acting on the interpretation of data.
Statistics is about uncertainty. It provides the mathematical approaches and tools to deal with uncertainty in a dataset of any size.
Statistics can be a powerful tool when performing the art of Data Science (DS). From a high-level view, statistics is the use of mathematics to perform technical analysis of data. Using statistics, we can gain deeper and more fine grained insights into how exactly our data is structured and based on that structure how we can optimally apply other data science techniques to get even more information.

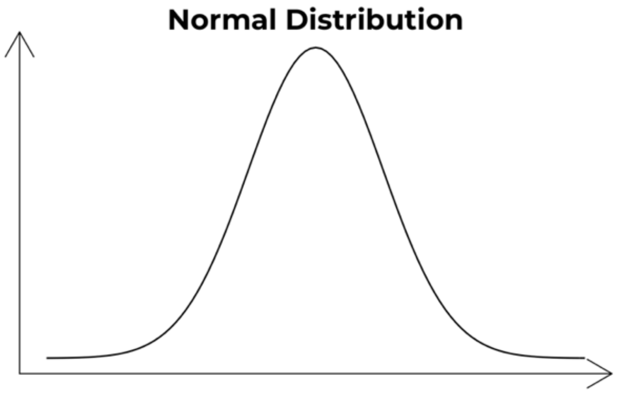
Descriptive/Summary Statistics

Summary statistics define a complicated set of data (or whole population) with some simple metrics. Basically, summary statistics summarise large amounts of data by describing key characteristics such as the average, distribution, potential correlation or dependence, etc. They are calculated using quantitative data that can be discrete or continuous, but they are not used when working with qualitative data.
Summary statistics usually fall into several broad categories: location, shape, spread, dependence, and order statistics. Summary statistics measure:

- Average (or central tendency) - Where is the data centered? Where is the trend? Examples include mode, median, and mean.
- Shape - How is the data distributed? What is the pattern? How is the data skewed? Examples include skewness or kurtosis and L-moments.
- Spread - How varied or dispersed is the data? Examples include range, variance, and standard deviation (among others).
- Dependence - If the data contains more than one variable, are the variables correlated or dependent? The primary example is correlation coefficient.

To focus on the shape, we can look at distribution which is a function that shows the possible values for a variable and how often they occur. The distribution of data is the shape of the data, gathered from an understanding of its centre and spread. A probability distribution gives the likelihood of obtaining each possible value. Probability distribution functions always add up to 1 since they are the sum of all possible values.
There are a number of common distributions that appear regularly in real-world scenarios, these can be both discrete and continuous.

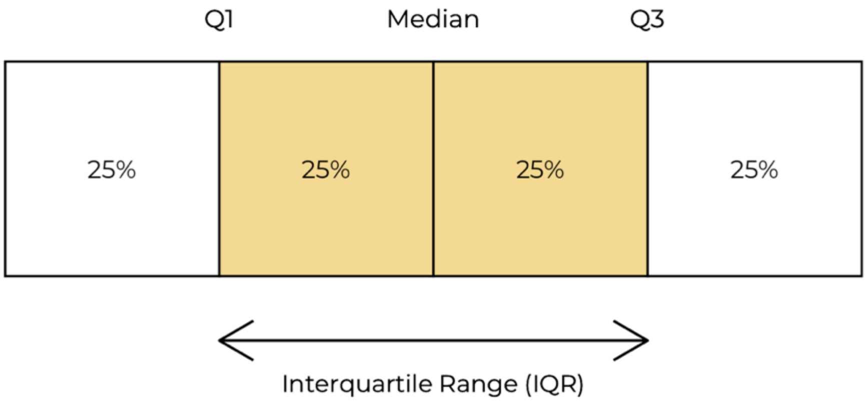
The normal distribution, also called the Gaussian distribution or a bell-curve is the most important distribution, as it occurs most commonly in nature. It also has some special properties, as its mean, mode and median are all equal.



Real-life examples of things that follow normal distributions are:

- Adult heights
- Blood pressure
- IQ
- Test scores

When looking at dispersion, this relates to understanding how data is spread, or dispersed, around the middle value. This gives a feel for how well the mean and median can summarise the data. In a similar way to the mean, the range is also very affected by outliers. The range is the largest (maximum) value minus the smallest (minimum) value in a dataset variable. To address this in a similar way to using the median to estimate the middle of the dataset rather than the mean, the interquartile range gives a view of the spread of the middle 50% of the data and is therefore unaffected by extreme values. The Interquartile range (IQR): upper quartile (Q3) minus the lower quartile (Q1) of a dataset variable.



Inferential Statistics

With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

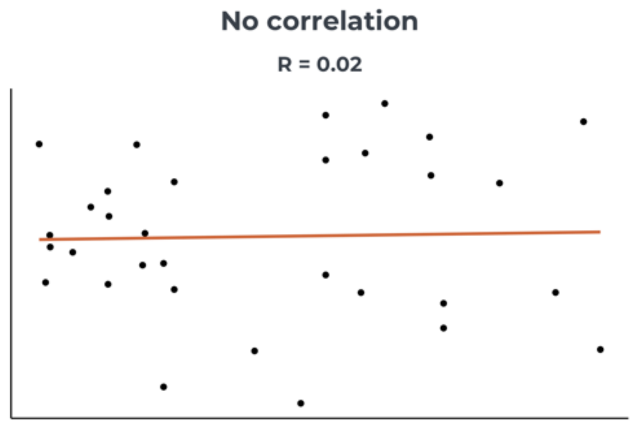
Regression and correlation analysis are statistical techniques used extensively to examine causal relationships between variables. Regression and correlation measure the degree of relationship between two or more variables in two different but related ways.

Correlation

A simple definition of correlation is the relationship between two or more variables (or data sets). This relationship is defined more specifically by its strength and direction.

What is a strong correlation in statistics?

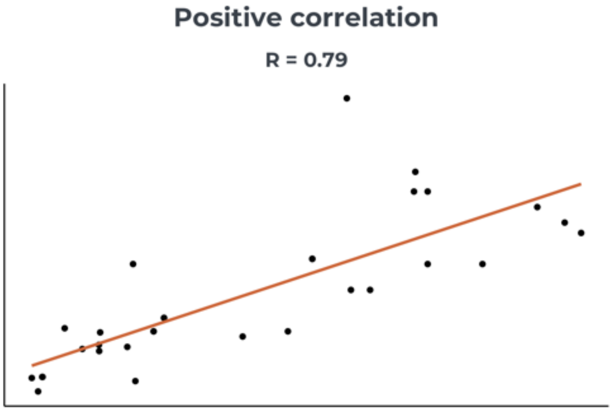
A strong correlation (sometimes referred to as high correlation) is when two groups of data are very closely related. The inverse is also true - weak (or low) correlation means the two groups of data are only somewhat related.For example, increasing ice cream sales have a strong (high) correlation with rising temperatures. The hotter it is, the more ice cream people eat.

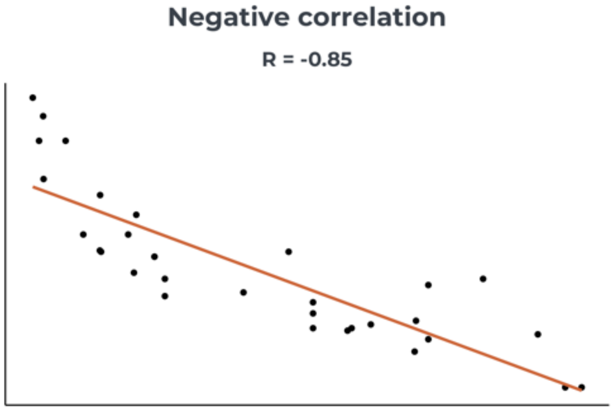


What are positive and negative correlations?

The direction of related variables can be either positive or negative. Positive correlation means both values increase together. Negative correlation means one value increases while the other value decreases. Continuing with the ice cream example, higher ice cream sales have a strong positive correlation with warmer temperatures because as the weather gets hotter (increasing value) more ice cream is sold (increasing value).

One negative correlation might be that less hot chocolate is sold (decreasing value) when the temperature gets warmer (increasing value). Or consider another example of negative correlation - the more someone pays on their mortgage (increasing value), the less they owe (decreasing value).





Regression

What is regression in statistics?

Regression in statistics is a useful way of making predictions about data. Regression shows the most likely outcome based on a trend of one or more known data points (predictors) - and the impact of changing one of the predictors.

For example, we might predict the most likely grade point average (GPA) students would earn at university based on their annual aptitude scores in high school.

The more technical definition of regression is the strength of the relationship between one or more independent data points (variables we can change or predictors) and one dependent data point (the predicted outcome).

Continuing with the example above, regression allows us to estimate how much higher students' college grades might be if their aptitude scores were improved by 2 points every year.

Regression example

Let's look at another regression example. Suppose you want to predict how much money you could make by investing in mutual funds over the next 10 years. The known data points (predictors or variables you can change) in this example would be how much money you contribute, how frequently you contribute, and the past performance of the mutual funds. By adjusting any one of those variables, you can predict how your return on investment may increase or decrease.

Regression equation and regression line

Regression equation and regression line are two important terms to know. A regression line is the trend that emerges when we plot our known data (predictors) on a graph. And the way we plot our data is by using a regression equation - a mathematical formula where we can plug in our known data to calculate the predicted outcome. There are different types of regression equations, but the most common one is the linear regression equation. (Learn more about regression equations here.)

By using a regression equation to visualise our data on a graph, we can more easily see how the outcome might change when one or more predictors change.

Types of regression

For different types of data, there are different types of regression. (Defining the data types in the context of regressions exceeds the depth of this explanation, but you can learn more about both types here.)

The types of regression include linear, logistic, polynomial, Bayesian, and others. At a very high level, the difference between regressions is the shape or arc of the regression line. Basically, each regression has a different shape when visualised on a graph (reflecting the varied patterns of the data being visualised). (It's worth noting that each type of regression has its own equation.)

What is the regression fallacy?

The regression fallacy, more commonly called regression toward the mean, is when something happens that's unusually good or bad reverts back towards the average (i.e. regresses toward the mean). This statistical fallacy occurs anywhere random chance plays a part in the outcome.

For example, success in business is often a combination of both skill and luck. This means that the best-performing companies today are likely to be much closer to average in 10 years time, not through incompetence but because today they're likely benefitting from a string of good luck – like rolling a double-six repeatedly.

Regression vs. correlation

Correlation shows what, if any, relationship exists between two data points.

Regression involves causation where one piece of information (outcome) is the effect of one or more other data points. Also, regression allows us to 'play' with the outcome by changing the independent data.

For example, we could see how fluctuating oil costs would impact petrol prices.

References

Alharthi, H. 2018. Healthcare predictive analytics: An overview with a focus on Saudi Arabia. Journal of Infection and Public Health. 11, 749 - 756.
Bruce, P. and Bruce A. 2017. Practical Statistics for Data Science. California: O'Reilly.
Costa, C. 2020. Best Data Science Tools for Data Scientists. Towards Data Science.
https://towardsdatascience.com/best-data-science-tools-for-data-scientists-75be64144a88
Geckoboard. Data Science Glossary. Available:
https://www.geckoboard.com/best-practice/data-science-glossary/#data-analytics
Ghosh, P. 2017. Fundamentals of Descriptive Analytics. Dataversity. Available at:
https://www.dataversity.net/fundamentals-descriptive-analytics/
Lowndes, J., Best, B., Scarborough, C. et al. Our path to better science in less time using open data science tools. Nat Ecol Evol 1, 0160 (2017).
https://doi.org/10.1038/s41559-017-0160
Trochim, W. Inferential Statistics. Research Methods Knowledge Base. Available:
https://conjointly.com/kb/inferential-statistics/