Analyzing Data

Data Analysis Definition

Data analysis is the process of mathematically summarizing data and evaluating patterns in data with the goals of discovering useful information, informing conclusions, and supporting decision making.

Five Types of Data Analysis

Different types of data analysis are needed for data from different sources and support different types of conclusions.
The five main types of data analysis are:

1. Descriptive analysis
2. Exploratory analysis
3. Inferential analysis
4. Causal analysis
5. Predictive analysis

Descriptive Analysis

In descriptive analyses, we calculate measures of central tendency and spread to summarize major patterns in a dataset.
Examples of measures of central tendency include: mean, median, mode.
Examples of measures of spread include: range, interquartile range, standard deviation, variance
Descriptive analysis also often include plots that help visualize measures of central tendency and spread. Common examples are box plots and histograms.

Limits of Descriptive Analysis

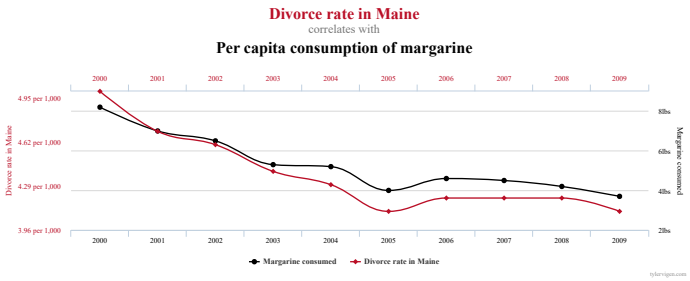
One limit of descriptive analysis is that the conclusions we draw cannot be extended beyond the data we directly analyzed.
For example, if we do a descriptive analysis on a dataset of household water usage in one region, we might find that the mean water usage is increasing over time. However, we would not be able to conclude anything about the mean water usage in other regions.

Exploratory Analysis

Exploratory data analysis looks for relationships between variables within a dataset. Exploratory analyses might reveal correlations between variables or group subsets of data based on shared characteristics.

Correlation and Causation

Correlation between variables does not necessarily mean a causal relationship exists between those variables.
For example, divorce rate in Maine and margarine consumption are correlated but margarine consumption does not cause divorces and divorce does not cause margarine consumption.



Inferential Analysis

Inferential analysis lets us draw conclusions about an entire population based on results from a subset or sample of that population. A/B testing, where we test which online feature performs better with a sample of a population, is a popular business application of inferential analysis.

Requirements for Inferential Analysis

Inferential analysis is a powerful tool. As a result, several rules need to be followed for the analysis to be valid:

1. The sample selected must be “big enough” in comparison to the population. 10% is a good rule-of-thumb.
2. The sample should be randomly selected and representative of the total population.
3. Only test one hypothesis at a time. Manipulating more than one variable makes it impossible to tell which variable influenced the outcome.

Causal Analysis

Causal analysis coupled with careful experimental design lets us go beyond correlation and actually assign causation.
Key factors of good experimental design are:

1. Control: only one variable is changed at a time and the rest are kept from influencing the outcome of the experiment.
2. Randomization: subjects are randomly selected and randomly assigned to treatment groups.
3. Replication: many subjects are included in the experiment and the experiment is repeated with the same results.

Causal Analysis with Observational Data

Sometimes we need to know why something happened but we cannot perform the necessary experiments because they are too expensive, unethical, or otherwise impossible. In such cases, we may be able to do causal analysis on observational data but it requires meeting strict assumptions and applying advanced techniques.
For example, climate scientists apply advanced causal analysis techniques to determine whether global climate change impacts local weather systems since planet-scale experiments are impossible.

Predictive analysis

Predictive analysis takes advantage of supervised machine learning techniques to estimate the likelihood of future outcomes.
For example, recommendation algorithms use the preferences of many other people together with your previous choices to predict what you are most likely to enjoy.

Supervised Machine Learning in Predictive Analysis

Examples of supervised machine learning techniques used in predictive analysis include: regression models, support vector machines, and convolutional neural networks.
Supervised machine learning is distinct from unsupervised machine learning because it always requires training data, or pre-labeled or classified data used to generate the predictive model.

Garbage In, Garbage Out

The quality of the predictions made during a predictive analysis is deeply dependent on the quality of the data used to generate the predictions.
For example, if a model is trained with mislabeled data, it will produce inaccurate predictions no matter how good the actual algorithm is. This is commonly referred to as, “garbage in, garbage out.”