Six Types Of Analyses Every Data Scientist Should Know

Jeffrey Leek, Assistant Professor of Biostatistics at John Hopkins Bloomberg School of Public Health, has identified six (6) archetypical analyses. As presented, they range from the least to most complex, in terms of knowledge, costs, and time. In summary, Descriptive, Exploratory, Inferential, Predictive, Causal and Mechanistic.

1. Descriptive (least amount of effort): The discipline of quantitatively describing the main features of a collection of data. In essence, it describes a set of data.

– Typically the first kind of data analysis performed on a data set
– Commonly applied to large volumes of data, such as census data
– The description and interpretation processes are different steps
– Univariate and Bivariate are two types of statistical descriptive analyses.

– *Type of data set applied to:* Census Data Set
– a whole population

2. Exploratory: An approach to analyzing data sets to find previously unknown relationships.

– Exploratory models are good for discovering new connections
– They are also useful for defining future studies/questions
– Exploratory analyses are usually not the definitive answer to the question at hand, but only the start
– Exploratory analyses alone should not be used for generalizing and/or predicting
– Remember: correlation does not imply causation

– *Type of data set applied to:* Census and Convenience Sample Data Set (typically non-uniform)
Example:– a random sample with many variables measured

3. Inferential: Aims to test theories about the nature of the world in general (or some part of it) based on samples of "subjects" taken from the world (or some part of it). That is, use a relatively small sample of data to say something about a bigger population.

– Inference is commonly the goal of statistical models
– Inference involves estimating both the quantity you care about and your uncertainty about your estimate
– Inference depends heavily on both the population and the sampling scheme

– *Type of data set applied to:* Observational, Cross Sectional Time Study, and Retrospective Data Set
Example: randomly sampled population

4. Predictive: The various types of methods that analyze current and historical facts to make predictions about future events. In essence, to use the data on some objects to predict values for another object.

– The models predicts, but it does not mean that the independent variables cause
– Accurate prediction depends heavily on measuring the right variables
– Although there are better and worse prediction models, more data and a simple model works really well
– Prediction is very hard, especially about the future references

– *Type of data set applied to:* Prediction Study Data Set
Example: a training and test data set from the same population

5. Causal: To find out what happens to one variable when you change another.

– Implementation usually requires randomized studies
– There are approaches to inferring causation in non-randomized studies
– Causal models are said to be the "gold standard" for data analysis

– *Type of data set applied to:* Randomized Trial Data Set
Example: data from a randomized study

6. Mechanistic (most amount of effort): Understand the exact changes in variables that lead to changes in other variables for individual objects.

– Incredibly hard to infer, except in simple situations
– Usually modeled by a deterministic set of equations (physical/engineering science)
– Generally the random component of the data is measurement error
– If the equations are known but the parameters are not, they may be inferred with data analysis

– *Type of data set applied to:* Randomized Trial Data Set
Example: data about all components of the system

Jerry Smith: https://datascientistinsights.com/2013/01/29/six-types-of-analyses-every-data-scientist-should-know/