Notes on Data Science questions

ME

12/2/2019

## Types of questions in Data Science —–

There are, broadly speaking, six categories in which data analyses fall. In the approximate order of difficulty, they are:

Descriptive - The goal of descriptive analysis is to describe or summarize a set of data. Whenever you get a new dataset to examine, this is usually the first kind of analysis you will perform. Descriptive analysis will generate simple summaries about the samples and their measurements. You may be familiar with common descriptive statistics: measures of central tendency (eg: mean, median, mode) or measures of variability (eg: range, standard deviations or variance).

Exploratory - The goal of exploratory analysis is to examine or explore the data and find relationships that weren’t previously known. Exploratory analyses explore how different measures might be related to each other but do not confirm that relationship as causitive. You’ve probably heard the phrase “Correlation does not imply causation” and exploratory analyses lie at the root of this saying. Just because you observe a relationship between two variables during exploratory analysis, it does not mean that one necessarily causes the other.

Inferential - The goal of inferential analyses is to use a relatively small sample of data to infer or say something about the population at large. Inferential analysis is commonly the goal of statistical modelling, where you have a small amount of information to extrapolate and generalize that information to a larger group.

Predictive - The goal of predictive analysis is to use current data to make predictions about future data. Essentially, you are using current and historical data to find patterns and predict the likelihood of future outcomes.

Causal - The caveat to a lot of the analyses we’ve looked at so far is that we can only see correlations and can’t get at the cause of the relationships we observe. Causal analysis fills that gap; the goal of causal analysis is to see what happens to one variable when we manipulate another variable - looking at the cause and effect of a relationship.

Mechanistic - Mechanistic analyses are not nearly as commonly used as the previous analyses - the goal of mechanistic analysis is to understand the exact changes in variables that lead to exact changes in other variables. These analyses are exceedingly hard to use to infer much, except in simple situations or in those that are nicely modeled by deterministic equations. Given this description, it might be clear to see how mechanistic analyses are most commonly applied to physical or engineering sciences; biological sciences, for example, are far too noisy of data sets to use mechanistic analysis. Often, when these analyses are applied, the only noise in the data is measurement error, which can be accounted for.

## Experimenal study design —–

When designing your experiment, it is important to balance your counfounders (randomization) between your experimental group and your control to avoid systematic errors

Replicability reduces the variability and the risk of obtaining results by chance P-value

P-hacking - means manipulating your data to achieve statistically significant results that have not real causality or relationship

hypothesis, essentially an educated guess as to the relationship between your variables and the outcome of your experiment.

Independent variable (AKA factor): The variable that the experimenter manipulates; it does not depend on other variables being measured. Often displayed on the x-axis.

Dependent variable: The variable that is expected to change as a result of changes in the independent variable. Often displayed on the y-axis, so that changes in X, the independent variable, effect changes in Y.

# Big data —–

Collection can be big on multiple domains - velocity - volume - variety - unstructure