The Data Scientists Toolbox - Week 3 Quiz

Statistical Inference

- In statistics, statistical inference is the process of drawing conclusions from data that are subject to random variation, for example, observational errors or sampling variation.
- Initial requirements of such a system of procedures for inference and induction are that the system should produce reasonable answers when applied to well-defined situations and that it should be general enough to be applied across a range of situations.
- Inferential statistics are used to test hypotheses and make estimations using sample data. Whereas descriptive statistics describe a sample, inferential statistics infer predictions about a larger population that the sample represents.
- The outcome of statistical inference may be an answer to the question "what should be done next?", where this might be a decision about making further experiments or surveys, or about drawing a conclusion before implementing some organizational or governmental policy.

Exploratory Data Analysis

- In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.
- Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.

Question 1 - Types of Data Analysis

We take a random sample of individuals in a population and identify whether they smoke and if they have cancer. We observe that there is a strong relationship between whether a person in the sample smoked or not and whether they have lung cancer.

We claim that the smoking is related to lung cancer in the larger population. We explain we think that the reason for this relationship is because cigarette smoke contains known carcinogens such as arsenic and benzene, which make cells in the lungs become cancerous.

(i) This is an example of an inferential data analysis.

(ii) This is an example of a mechanistic data analysis.

(iii) This is an example of an descriptive data analysis.

(iv) This is an example of a causal data analysis.

Question 2

What is the most important thing in Data Science?

(i) Working with large data sets.

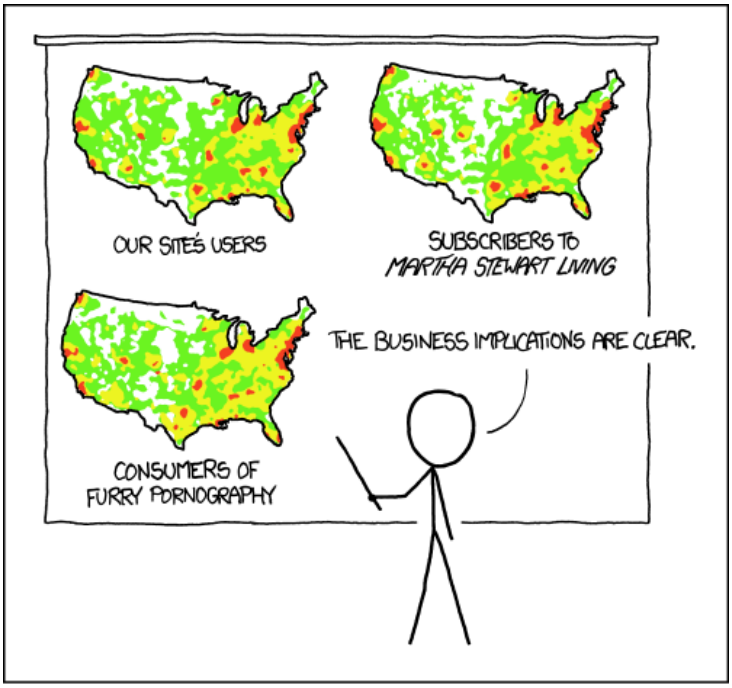
(ii) Hacking skills.

(iii) Knowing Hadoop and Pig.

(iv) The question you are trying to answer.

Question 3

If the goal of a study was to relate *Martha Stewart Living Subscribers* to *Our Site's Users* based on the number of people that lived in each region of the US, what would be the potential problem?



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

(i) There would be confounding because the number of people that live in an area is related to both *Martha Stewart Living Subscribers* and *Our Site's Users*.

(ii) We would be performing inference on the relationship between *Martha Stewart Living Subscribers* and *Our Site's Users*.

(iii) We wouldn't know the sensitivity of our predictions.

(iv) We wouldn't be able to estimate the variability in *Martha Stewart Living Subscribers*.

Experimental Design

- Experimental design is the design of any information-gathering exercises where variation is present, whether under the full control of the experimenter or not.
- In the design of experiments, the experimenter is often interested in the effect of some process or intervention (the "treatment") on some objects (the "experimental units"), which may be people, parts of people, groups of people, plants, animals, etc.
- Design of experiments is thus a discipline that has very broad application across all the natural and social sciences and engineering.

- We are concerned with the analysis of data generated from an experiment. It is wise to take time and effort to organise the experiment properly to ensure that the right type of data, and enough of it, is available to answer the questions of interest as clearly and efficiently as possible. This process is called experimental design.

- The specific questions that the experiment is intended to answer must be clearly identified before carrying out the experiment.

- We should also attempt to identify known or expected sources of variability in the experimental units since one of the main aims of a designed experiment is to reduce the effect of these sources of variability on the answers to questions of interest. That is, we design the experiment in order to improve the precision of our answers.

Treatment

- In experiments, a treatment is something that researchers administer to experimantal units .

- For example, a corn field is divided into four, each part is 'treated' with a different fertiliser to see which produces the most corn; a teacher practices different teaching methods on different groups in her class to see which yields the best results; a doctor treats a patient with a skin condition with different creams to see which is most effective.

- Treatments are administered to experimental units by 'level', where level implies amount or magnitude. For example, if the experimental units were given 5mg, 10mg, 15mg of a medication, those amounts would be three levels of the treatment.

- 'Level' is also used for categorical variables, such as Drugs A, B, and C, where the three are different kinds of drug, not different amounts of the same thing.

Factor

- A factor of an experiment is a controlled independent variable; a variable whose levels are set by the experimenter.

- A factor is a general type or category of treatments. Different treatments constitute different levels of a factor. For example, three different groups of runners are subjected to different training methods. The runners are the experimental units, the training methods, the treatments, where the three types of training methods constitute three levels of the factor 'type of training'.

Question 4

What is an experimental design tool that can be used to address variables that may be confounders at the design phase of an experiment?

(i) Randomization

(ii) Only using non-confounding variables.

(iii) Using data from a database.

(iv) Using all the data you have access too.

Question 5

What is the reason behind the explosion of interest in big data?

(i) We recently discovered ways to use data to make predictions.

(ii) The price and difficulty of collecting and storing data has dramatically dropped.

(iii) We recently discovered ways to use data to answer scientific and business questions.

(iv) There have been massive improvements in machine learning algorithms.