

The Data Scientists Toolbox - Week 3 Quiz

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.

- This is an example of an inferential data analysis.
- This is an example of a mechanistic data analysis.
- This is an example of an descriptive data analysis.
- 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?

Link to image Source: <http://xkcd.com/1138/>

- (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*.

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.