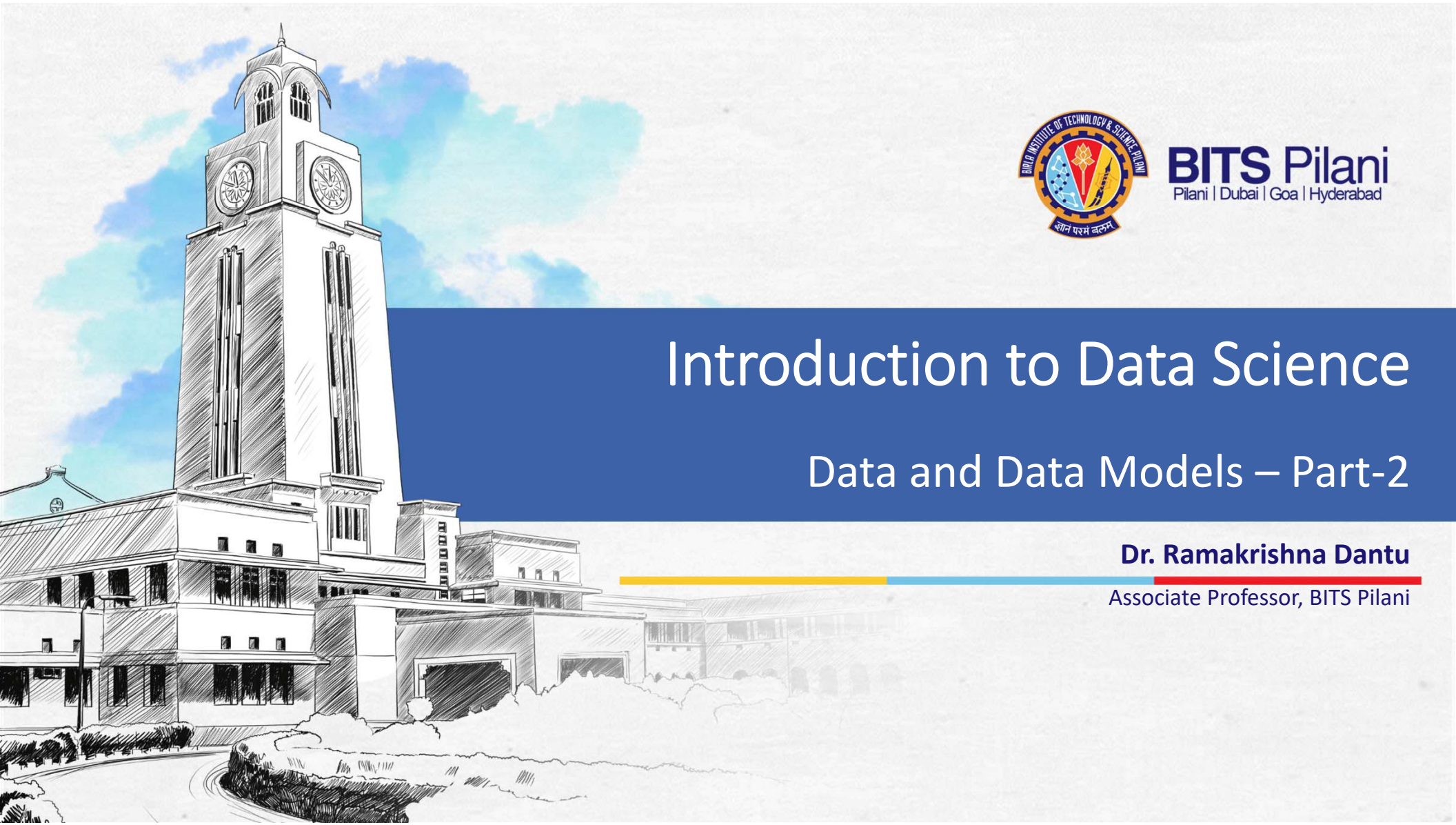


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Data and Data Models – Part-2

Dr. Ramakrishna Dantu
Associate Professor, BITS Pilani

innovate achieve lead

Disclaimer and Acknowledgement

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Introduction to Data Science

innovate achieve lead

Data and Data Models

- Types of Data and Datasets
- Data Quality
- Epicycles of Data Analysis
- Data Models
  - Model as expectation
  - Comparing models to reality
  - Reactions to Data
  - Refining our expectations
- Six Types of the Questions
- Characteristics of Good Question
- Formal modelling
  - General Framework
  - Associational Analyses
  - Prediction Analyses

Types of Research Questions

Research Questions

Six Type of Research Questions

- The type of question we ask directly informs how we interpret our results
  - Descriptive
  - Exploratory
  - Inferential
  - Predictive
  - Causal
  - Mechanistic

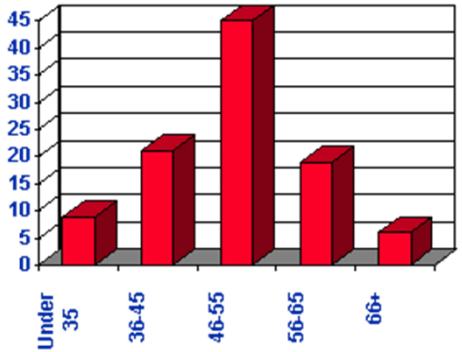
Source: The Art of Data Science by Peng & Matsui

Research Questions

innovate achieve lead

Descriptive Question

- A descriptive question seeks to summarize a characteristic of a set of data
- Examples include:
  - determining the proportion of males vs females who smoke
  - the mean number of servings of fresh fruits and vegetables per day
  - the frequency of viral illnesses in a set of data collected from a group of individuals
- There is no interpretation of the result, as the result is a fact, an attribute of the set of data
  - We only describe the data in general language



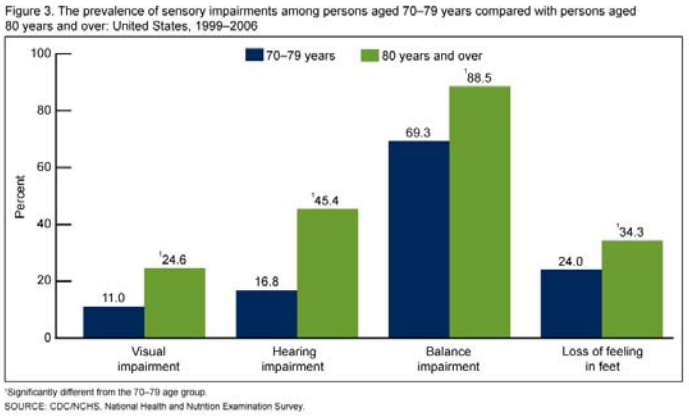
Category	Percent
Under 35 years old	9%
36–45	21%
46–55	45%
56–65	19%
66+	6%

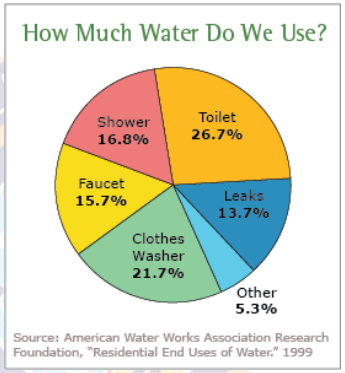
Source: The Art of Data Science by Peng & Matsui

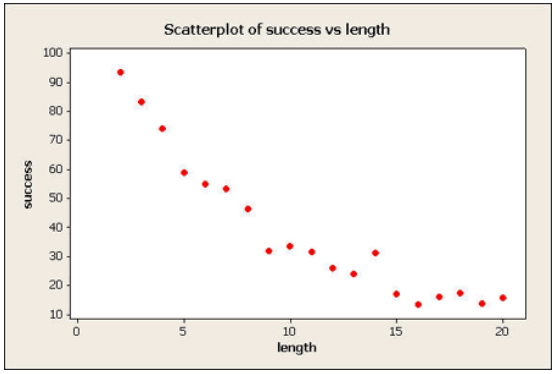
Research Questions

innovate achieve lead

Descriptive Question - Examples







Example: Nutrient Intake Data – Descriptive Statistics

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Calcium	737	624.0492537	397.2775401	7.4400000	2866.44
Iron	737	11.1298996	5.9841905	0	58.6680000
Protein	737	65.8034410	30.5757564	0	251.0120000
A	737	839.3653460	1633.54	0	34434.27
C	737	78.9284464	73.5952721	0	433.3390000

Source: https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/

Research Questions

innovate achieve lead

Exploratory Question

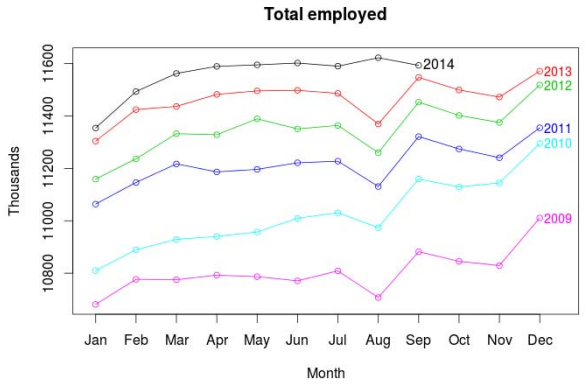
- If the question is exploratory, then we analyze the data to see if there are patterns, trends, or relationships between variables
- These types of analyses are also called "hypothesis-generating" analyses
  - because rather than testing a hypothesis as would be done with an inferential, causal, or mechanistic question, we are looking for patterns that would support proposing a hypothesis
- Imagine that we have a general idea that diet was linked somehow to viral illnesses
  - Here, we might explore this idea by examining relationships between a range of dietary factors and viral illnesses
- For example:
  - If we find in our analysis that individuals who ate a diet high in certain foods had fewer viral illnesses than those whose diet was not enriched for these foods
  - Then, we propose the hypothesis that among adults, eating at least 5 servings a day of fresh fruit and vegetables is associated with fewer viral illnesses per year

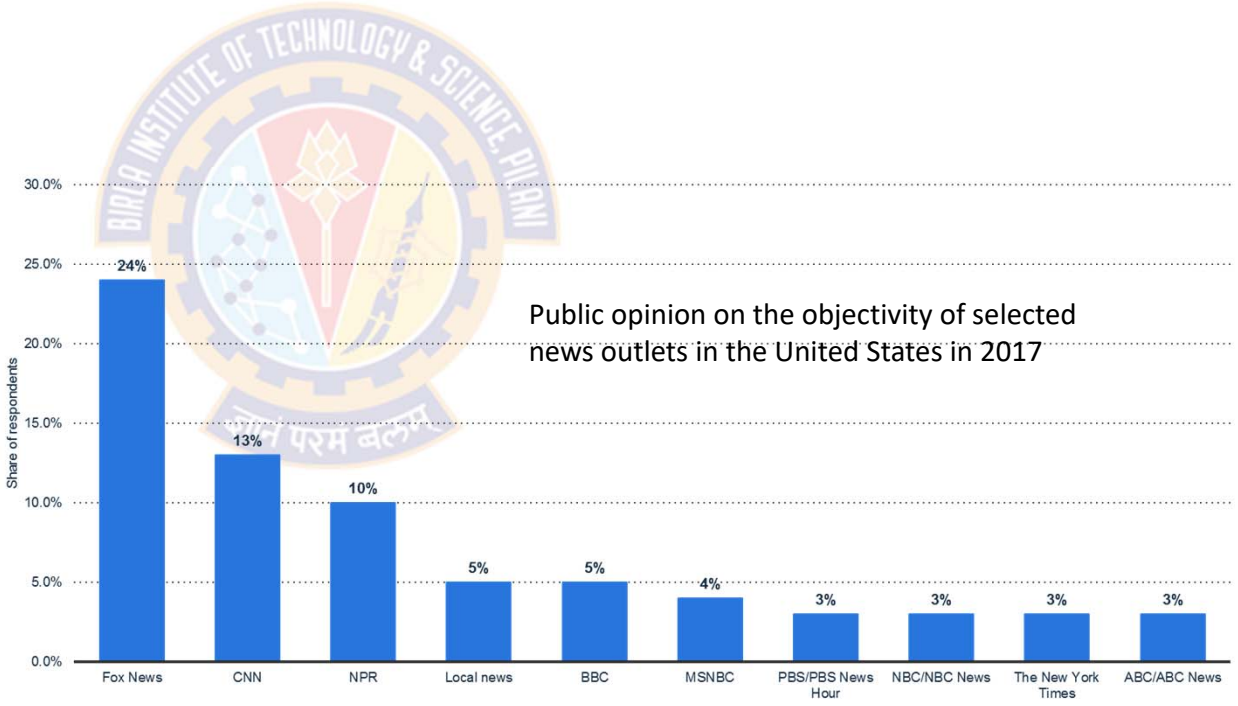
Source: The Art of Data Science by Peng & Matsui

Research Questions

innovate achieve lead

Exploratory Question - Examples



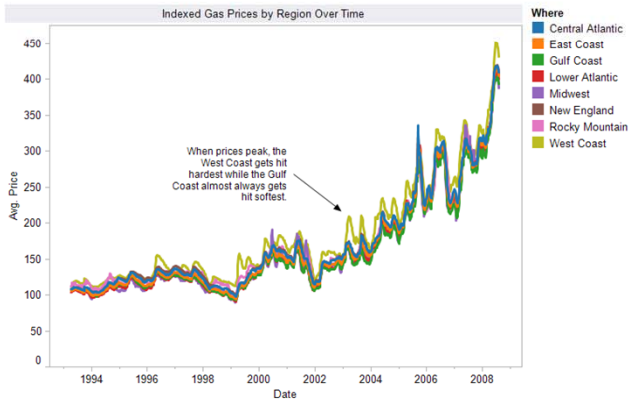


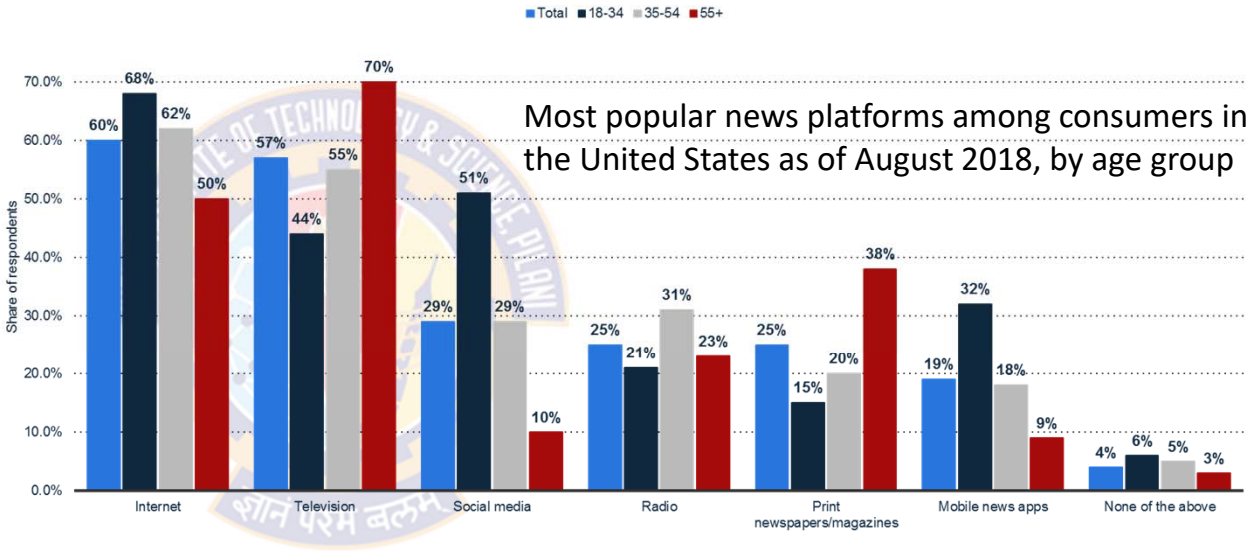
Public opinion on the objectivity of selected news outlets in the United States in 2017

Research Questions

innovate achieve lead

Exploratory Question - Examples





Most popular news platforms among consumers in the United States as of August 2018, by age group

Research Questions

Inferential Question

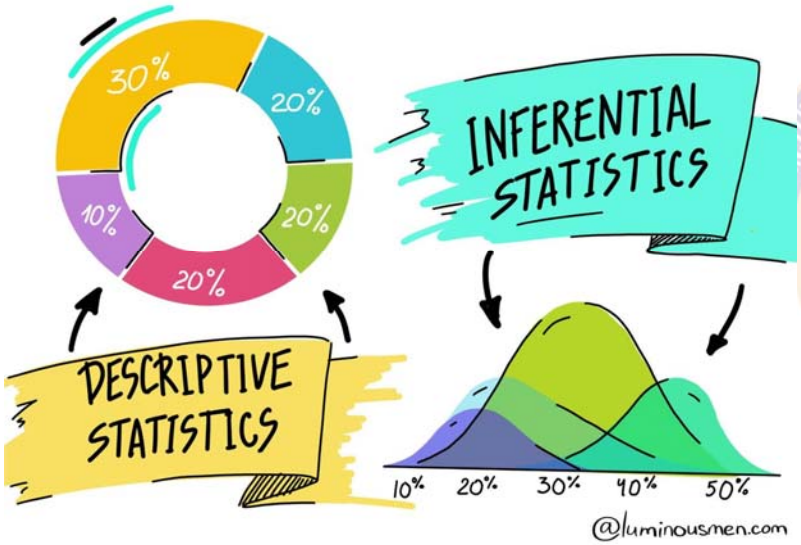
- An inferential question would be a restatement of this proposed hypothesis as a question
- Involves methods for drawing and measuring the reliability of conclusions about a population based on the information obtained from a sample of the population
- This question would be answered by analyzing a different set of data
  - could be a representative sample of adults in India
- By analyzing this different set of data we accomplish two things:
  - determine if the association we observed in our exploratory analysis holds in a different sample
  - determine whether it holds in a sample that is representative of the adult Indian population
    - this suggests that the association is applicable to all adults in India
- In other words, we will be able to infer what is true, on average, for the adult population in India from the analysis we perform on the representative sample

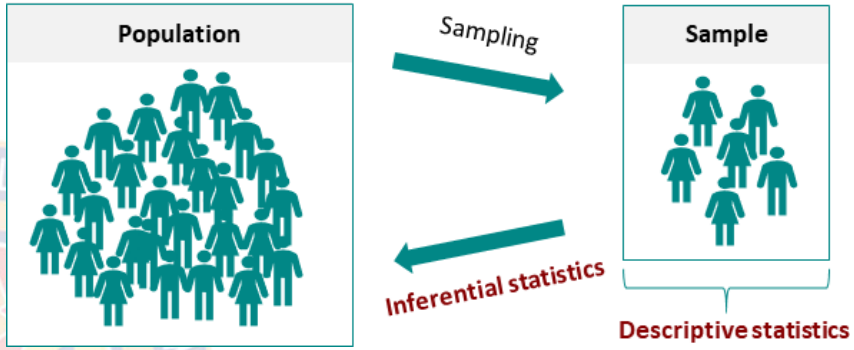
Source: The Art of Data Science by Peng & Matsui

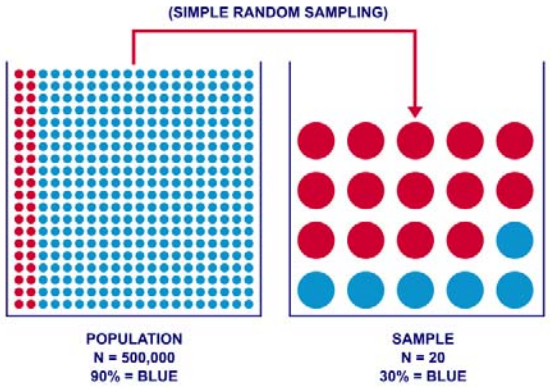
Research Questions

innovate achieve lead

Inferential Question - Examples







Source: The Art of Data Science by Peng & Matsui

Research Questions

innovate achieve lead

Predictive Question

- A predictive question is the one where we ask what types of people will eat a diet high in fresh fruits and vegetables during the next year
- In this type of question you are less interested in what causes someone to eat a certain diet
  - We are interested in knowing what makes someone *will* eat a certain diet
- For example, higher income may be one of the final set of predictors
- We may not be interested in why people with higher incomes are more likely to eat a diet high in fresh fruits and vegetables
- What is most important us is that income is a factor that predicts this behavior

Source: The Art of Data Science by Peng & Matsui

Research Questions

innovate achieve lead

Causal Question

- A causal question asks about whether changing one factor will change another factor, on average, in a population
- An inferential question might tell us that people who eat a certain type of foods tend to have fewer viral illnesses
- The answer to this question does not tell us if eating these foods causes a *reduction* in the number of viral illnesses
  - This is the case for a causal question

Source: The Art of Data Science by Peng & Matsui

Research Questions

innovate achieve lead

Causal Question

- Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal. For example,
  - Data collected in the context of a randomized trial
  - People are randomly assigned to eat a diet high in fresh fruits and vegetables or one that was low in fresh fruits and vegetables
- Even if our data are not from a randomized trial, we can take an analytic approach designed to answer a causal question

Source: The Art of Data Science by Peng & Matsui

Research Questions

innovate achieve lead

Causal Question

Research Purpose	Research Question	Hypothesis
To see if increasing service or support staff be profitable	What is the relationship of the size of service staff and revenue?	An increase in 25% of service staff results in marginal revenue higher than marginal cost
To understand which advertising campaign for public transit system should be run	What makes people get of their cars into public transit system	Advertising program A generates more new riders than program B
Should we introduce a "No Frills" class of airfare?	Will the "no frills" airfare generate sufficient passengers to offset the loss of revenue?	The new airfare will attract 25% more passengers to generate sufficient revenue

Source: The Art of Data Science by Peng & Matsui

Research Questions

innovate achieve lead

Mechanistic Question

- If the diet does indeed cause a reduction in the number of viral illnesses, *how* the diet leads to such a reduction?
- None of the questions described so far will not answer this "how" question
- A question that asks *how* a diet high in fresh fruits and vegetables leads to a reduction in the number of viral illnesses would be a *mechanistic* question

Source: The Art of Data Science by Peng & Matsui

Research Questions

innovate achieve lead

Notes

- First, many data analyses answer multiple types of questions
- If a data analysis aims to answer an inferential question, descriptive and exploratory questions must also be answered during the process of answering the inferential question
- For instance, in our example of diet and viral illnesses, we don't jump straight to a statistical model of the relationship between a diet high in fresh fruits and vegetables and the number of viral illnesses without having determined the frequency of this type of diet and viral illnesses and their relationship to one another in this sample
- A second point is that the type of question you ask is determined in part by the data available to you (unless you plan to conduct a study and collect the data needed to do the analysis)
- For example, to know whether eating a diet high in fresh fruits and vegetables causes a decrease in the number of viral illnesses, we would need before and after data:
  - People's diets change from one that is high in fresh fruits and vegetables to one that is not, or vice versa
- If this type of data set does not exist, then the best you may be able to do is either apply causal analysis methods to observational data or instead answer an inferential question about diet and viral illnesses.

Source: The Art of Data Science by Peng & Matsui

Characteristics of a Good Question

Characteristics of a Good Question

Introduction

- Five key characteristics of a good question for a data analysis:
  - Question should be interesting
  - Question should not already been answered
  - Question should be plausible
  - Question should be answerable
  - Question should be specific

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
ज्ञानं परमं बलम्

Source: The Art of Data Science by Peng & Matsui

Characteristics of a Good Question

innovate achieve lead

Question should be interesting

- At a minimum, the question should be interesting to the audience
- This depends on the context and environment in which you are working with data
- If you are in academia, the audience may be your collaborators, the scientific community, government regulators, your funders, and/or the public
- If you are working at a startup, your audience is your boss, the company leadership, and the investors
- For example:
  - The question of whether outdoor particulate matter pollution is associated with developmental problems in children may be of interest to people involved in regulating air pollution, but may not be of interest to a grocery store chain
  - Similarly, the question of whether sales of a detergent powder are higher when it is displayed with other dish washing items would be of interest to a grocery store manager, but not to others

Source: The Art of Data Science by Peng & Matsui

Characteristics of a Good Question

innovate achieve lead

Question should not already been answered

- The question is interesting, but if it's already answered, then it's of no use
- With the *explosion of data*, the growing amount of *publicly available data*, and the seemingly *endless scientific literature,* and other resources, it is possible that our question of interest has been answered already
- Background work and discussion with experts can help sort this out
  - Here, a brief review of extant literature would be required
- Even if the our research question has not been answered, related questions may have been answered
  - Answers to these related questions help in deciding if or how we can proceed with our specific question

Source: The Art of Data Science by Peng & Matsui

Characteristics of a Good Question

innovate achieve lead

Question should be plausible

- The question should also stem from a plausible (believable, convincing, credible) framework
- The question of whether the sales of detergent powder is correlated with its placement next to dishwashing items is a plausible one
  - because shoppers buying detergent are more likely than other shoppers to be interested in detergents
  - However, the question of whether detergent powder sales correlate with yogurt sales may be less plausible
    - unless you had some prior knowledge suggesting that these should be correlated.

Source: The Art of Data Science by Peng & Matsui

Characteristics of a Good Question

innovate achieve lead

Question should be plausible

- If you ask a question whose framework is not plausible, you are likely to end up with an answer that's difficult to interpret or have low confidence in
- In the detergent powder-yogurt question, if you do find they are correlated, many questions are raised about the result itself:
  - is it really correct?
  - why are these things correlated- is there another explanation?
  - is the data collected properly?
  - whether the data collected is really related to the question being asked?
- You can ensure that your question is grounded in a plausible framework (theory) by using your own knowledge of the subject area and doing a little research
- Our own knowledge of the subject area and research (literature review) together can help us in sorting out whether our question is grounded in a plausible framework (theory).

Source: The Art of Data Science by Peng & Matsui

Characteristics of a Good Question

innovate achieve lead

Question should be answerable

- The question should of course be answerable
- Some of the best questions aren't answerable, either because
  - the data don't exist or there is no means of collecting the data because of lack of resources, feasibility, or ethical problems
- For example,
  - If we want to study whether there are defects in the functioning of certain cells in the brain that cause autism
    - The question is quite plausible, but it not possible to perform brain biopsies to collect live cells to study, which would be needed to answer this question
  - If we want to investigate a research question about illness of certain category of patients where we need to survey a niche group of specialist medical doctors and the patients
    - Here, the question is plausible, but it is not answerable because it is not possible to get access to the population

Source: The Art of Data Science by Peng & Matsui

Characteristics of a Good Question

innovate achieve lead

Question should be specific

- An example of a general question is:
  - Is eating a healthier diet better for you?
- A more specific question emerges if we ask ourselves:
  - what do we mean by a “healthier” diet?
  - what is something is “better for you”?
- The process of question refinement leads to a specific question such as:
  - "Does eating at least 5 servings per day of fresh fruits and vegetables lead to fewer upper respiratory tract infections (colds)?"
- With this degree of specificity, our plan of attack is much clearer and the answer we will get at the end of the data analysis will be more interpretable
  - Because we either recommend or not recommend the specific action of eating at least 5 servings of fresh fruit and vegetables per day as a means of protecting against upper respiratory tract infections

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

Formal Modeling

innovate achieve lead

Introduction

- Formal modeling involves using mathematical notations for expressing our research models
- Often it is useful to represent a model using mathematical notation because:
  - it is a compact notation and
  - it can be easy to interpret once we get used to it
- Also, writing statistical model using mathematical notation, as opposed to just natural language, makes us to be precise:
  - in our description of the model and
  - in our statement of what we are trying to accomplish, such as estimating a parameter.

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

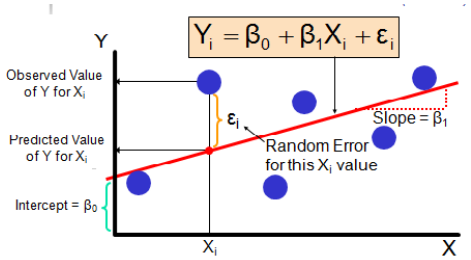
Goals of Formal Modeling

- One key goal of formal modeling is to develop a *precise specification of our question* and how our data can be used to answer that question
- Formal models allow us to identify clearly:
  - what we are trying to infer from data and
  - what form the relationships between features of the population take
- It can be difficult to achieve this kind of precision using words alone
- Parameters play an important role in many formal statistical models
  - in statistical language, these are known as parametric statistical models
- In formal models, we use numbers to represent features or associations existing in the population

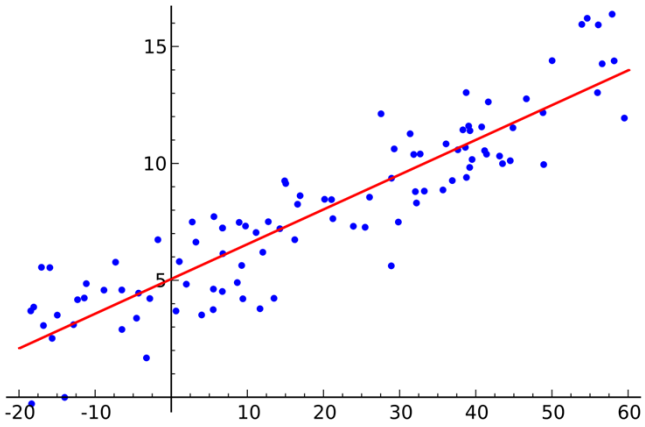
Source: The Art of Data Science by Peng & Matsui

Formal Modeling

Goals of Formal Modeling



- Parameters represent features in the population
  - Because of this they are generally considered unknown
- Our goal is to estimate them from the data we collect
- For example, suppose we want to assess the relationship between the number of ounces of soda consumed by a person per day and that person's BMI
- The slope of a line that we might plot visualizing this relationship is the parameter we want to estimate to answer your question:
  - "How much would BMI be expected to increase per each additional ounce of soda consumed?"
- More specifically, we are using a linear regression model to formulate this problem



Source: The Art of Data Science by Peng & Matsui

Formal Modeling

Goals of Formal Modeling

- Another goal of formal modeling is to develop a rigorous framework with which we can challenge and test our primary results
- At this point in our data analysis:
  - we've stated and refined our question,
  - we've explored the data visually and
  - maybe conducted some exploratory modeling
- We might likely have a pretty good sense of what the answer to our question is, but maybe have some doubts about whether our findings will hold up under intense scrutiny
- Assuming we are still interested in moving forward with our results, this is where formal modeling can play an important role

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

General Framework

- General framework of formal model involves three phases:
  - Setting Expectations
    - Initial expectations are set through the primary model
    - Primary model represents our best sense of what provides the answer to our research question
    - This model is chosen based on whatever information you have currently available.
  - Collecting Information
    - Once the primary model is set, we will want to create a set of secondary models that challenge the primary model in some way.
  - Revising Expectations
    - If our secondary models are successful in challenging our primary model and put the primary model's conclusions in some doubt, then we may need to adjust or modify the primary model to better reflect what we have learned from the secondary models

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

innovate achieve lead

Primary Model

- Primary model is usually derived from an exploratory data analyses
- Primary model succinctly summarizes the results and matches our expectations
- At any moment in a data analysis, the primary model is not necessarily the final model
  - It is simply the model against which we compare other secondary models
- The process of comparing the primary model to other secondary models is often referred to as sensitivity analyses
  - because we are interested in seeing how sensitive our model is to changes, such as adding or deleting predictors or removing outliers in the data
- Through the iterative process of formal modeling, we may decide that a different model is better suited as the primary model
- This is all part of the process of setting expectations, collecting information, and refining expectations based on the data

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

innovate achieve lead

Secondary Models

- Once we have decided on a primary model, we typically develop a series of secondary models
- The purpose of these models is to test the legitimacy and robustness of the primary model and potentially generate evidence against the primary model
- If the secondary models are successful in generating evidence that refutes the conclusions of the primary model,
  - then we may need to revisit the primary model and see whether its conclusions are still reasonable

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

innovate achieve lead

Association Analysis

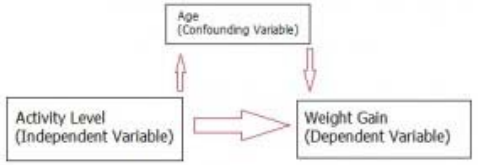
- In associational analyses, we study an association between two or more features in the presence of other potentially confounding factors
- Three classes of variables are important in an associational analysis:
  - Outcome variable
  - Predictor variable(s)
  - Potential confounders

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
ज्ञानं परमं बलम्

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

Association Analysis



- Outcome variable
  - The outcome is the feature of our dataset that is thought to change along with the *key predictor(s)*
- Key predictor
  - In associational analyses, there may be one or more key predictor(s) of interest
  - We want to know how the outcome changes with this key predictor
  - However, our understanding of that relationship may be challenged by the presence of potential confounders
- Potential confounders
  - This is a large class of predictors that are both related to the key predictor and the outcome
  - It's important to have a good understanding what these are and whether they are available in our dataset
  - If a key confounder is not available in the dataset, sometimes there will be a proxy that is related to that key confounder that can be substituted instead

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

innovate achieve lead

Association Analysis

- Once we have identified the three classes of variables in our dataset, we can begin formal modeling in an associational setting
- The basic form of a model in an associational analysis is:

$$y = \alpha + \beta x + \gamma z + \varepsilon$$

  - where
    - y is the outcome
    - x is the key predictor
    - z is a potential confounder
    - ε is independent random error
    - α is the intercept, i.e. the value y when x = 0 and z = 0
    - β is the change in y associated with a 1-unit increase x, controlling for z
    - γ is the change in y associated with a 1-unit increase in z, controlling for x

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

Association Analysis

$$y = \alpha + \beta x + \gamma z + \varepsilon$$

- This is a linear model, and our primary interest is in estimating the coefficient β, which quantifies the relationship between the key predictor x and the outcome y
- Even though we have to estimate α and γ as part of the process of estimating β, we do not really care about the values of those α and γ
- In the statistical literature, coefficients like α and γ are sometimes referred to as nuisance parameters because we have to use the data to estimate them to complete the model specification, but we do not actually care about their value

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

Prediction Analysis

- In the association analysis, the goal was to see if a key predictor x is associated with an outcome y
- But sometimes the goal is to use all of the information available to us to predict y
- In prediction models, we have outcome variables–features about which we would like to make predictions
  - but we typically do not make a distinction between "key predictors" and other predictors
- In most cases, any predictor that might be of use in predicting the outcome would be considered in the analysis
- Also, all predictors are given equal weight in terms of their importance in predicting the outcome
- Prediction analyses will often leave it in the prediction algorithm to determine the importance of each predictor and to determine the functional form of the model

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

Prediction Analysis

- In prediction analysis it is not possible to literally write down the model that is being used to predict because it cannot be represented using standard mathematical notation
- Many modern prediction routines are structured as algorithms or procedures that take inputs and transform them into outputs
- The path that the inputs take to be transformed into outputs may be highly nonlinear and predictors may interact with other predictors on the way
- Typically, there are no parameters of interest that we try to estimate
  - in fact many algorithmic procedures do not have any estimable parameters at all

Source: The Art of Data Science by Peng & Matsui

Formal Modeling

Prediction Analysis

- In prediction analyses, we usually do not care about the specific details of the model
- In most cases:
  - as long as the method "works",
  - is reproducible, and
  - produces good predictions with minimal error,

  then we have achieved our goals
- With prediction analyses, the precise type of analysis we do depends on the nature of the outcome
- Prediction problems typically come in the form of a classification problem where the outcome is binary


innovate
achieve
lead

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
ज्ञानं परमं बलम्

Thank You!