Workshop 5: Data, and How We Analyze It

Myles D. Garvey, Ph.D

Winter, 2020

1 Learning Objectives

1. To understand the differing goals of data analysis.
2. To understand how data is commonly structured.
3. To understand a unit-of-analysis as well as a "level" of analysis.
4. To understand the differences between a variable, construct, factor, and component.
5. To understand the difference between causation and association (i.e. correlation).
6. To understand the difference between cross-sectional, time-series, and panel data sets.
7. To understand the taxonomy of data analysis methods, and how they differ.
8. To understand the purpose of different data mining tools.

2 Workshop Description and Submission

You will submit to blackboard a .PDF document. Complete each exercise in this workshop. These are easy and small, and so it should not take you very long. In a word document, type up the question number, the original question text, and your response in R code. Save your word document as a .PDF and submit it to the submission link on Blackboard. Please note that failure to submit this as a .PDF will result in a 0 on the workshop.

3 Tutorial 1: Goals of Data Analysis

Put simply, data is a collection of observations. How those observations are organized, information-wise, is a different problem in and of itself. Putting this aside for the moment, we need to consider what different people need different data for. Data is everywhere, and is used by everyone. How people use this data, however, greatly differs based on their goals. Let us take something as simple as data on temperatures. To the meteorologist, historical daily temperature information may

help aid in the development of a mathematical model that can be used to predict the temperature over the course of the next few weeks. To the archaeologist, however, they may care less about prediction, and more about using the information as a way to *explain* other information such as bedrock formation. Needless to say, both individuals are using the same data, yet analyzing it differently.

Hence, one can ascertain that data analysis rests not necessarily on the data itself, but rather on the purpose for which the data will be used. This is not to say from where the data came is an irrelevant factor. On the contrary, analysts are almost always careful to fully understand the source of the data. However, how one approaches data analysis depends on both of these factors. Our example above illustrates two common goals for data analysis, scilicet *explanation* and *prediction*. Both will be explored in this workshop.

When it comes to explanation, the analyst's goal is to use the data to either describe characteristics about something, or, to validate a theory. A *theory* is a collection of *entities* (more on this later), characteristics of those entities, and the relationships between these characteristics. The theory is used to predict the outcome of characteristics when provided information about other characteristics. When we gather data for the goal of explaining a theory, were are trying to determine if the outcomes predicted by the theory are correct. Put simply, we are trying to determine if the theory is what actually explains the "thing" (whatever that may be) in the real world. Put differently, theory helps us explain the *how* and the *why* of what generated the data in the first place.

On the other hand, we may care strictly about prediction. In this instance, we do not care about a theory, nor do we care about that *how* or even the *why* of what generated that data in the first place. Our goal with prediction is simple: find a mathematical equation, which itself is found using data, which can accurately predict new observations. When we care about prediction, we do not care about how absurd or ridiculous the claim is. For example, if we happen to find that the total number of touch downs throughout all Sunday football games accurately predicts our sales levels for the week, then who cares how or why this happens? If we have an equation to use to plug in the total number of touch downs and get out the total amount of sales we will get this week, then it really doesn't matter if the two characteristics (namely touchdowns and sales) are actually related or totally spurious. If the prediction works, then it works, so who cares how or why it does?

Let us take a look at a few examples of work that where the goal of the data used in the analysis was for explanation or prediction. First, we will turn our attention to the paper by Lin et al. (2018), "Understanding olfaction and emotions and the moderating role of individual differences". Open the paper, and read the *abstract* to understand what it is about. Next, turn to page 820, and observe the Figure. As you can see, this paper is about a new theory for which has been proposed, and for which the data is being gathered for the goal of *validating* the theory. In this project, the theory is formed by using the existing literature on the matter, and is subsequently validated with newly observed data. Hence, in this project, the analyst's goal is to *explain* something (i.e. the theory) with the data, rather than to *predict* something with the data.

On the other hand, if we look at Kim & Han (2000), "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", we can see a different tone and structure in the analysis. If you read through the structure of the paper itself, you will not find a theoretical reasoning that is given. You will also not find a model that relates different characteristics to each other or a theoretical explanation as to why the characteristics

are related to each other. Instead, you merely find an algorithm and an equation that uses the data in what is called a "training set" to predict outcomes for a "test set", also called a "holdout sample". The study does not care about how or why the characteristics are related. Instead, all they care about is the design of an algorithm or equation which can use the data to accurately predict the outcomes of new observations. Hence, in this example, they are not using the data to *explain*, but rather to *predict*.

Summarizing this, we have two primary purposes of data: explanation and prediction. **Explanation** seeks to use data to better understand a concept, entity, group of entities, an entity's characteristic, or a relationship between various characteristics of entities. That is, when we use data to *explain*, we are trying to understand the *how* and the *why* of something. From the data analysis perspective, when we attempt to explain something with data, we usually are trying to *interpolate* the data. Put differently, we are trying to find an equation which can render an output, when given an input, that is as "close" to the value of an actual data point. For example, we may gather information on people's income and spending levels. We may use this data to find an equation to find someone's spending levels when provided information about that person's income levels. The way in which we use the data to find the equation is to minimize some objective (usually something called a mean square error (MSE)) which measures in some way how different (or similar) the value our equation gives for the spending when given a value of income from the actual spending of that individual. Put simply, interpolation compares "guesses" which are generated using all data points in our data against the "actual" values in our data.

On the other hand, when we are **predicting**, then the use of our data, and our analysis for it, greatly differs from that of explaining. When we try to use our data for prediction, we are instead trying to *extrapolate* rather than interpolate. Put simply, we are trying to use data to design an equation which can accurately predict *new* observations which were not used in the design of our equation. This differs from the goal of explanation, which uses an entire data set. Usually, when we are analyzing data for the purpose of prediction, we do not care about the "how" or the "why". We just simply care about the design of an equation which can predict new observations very well.

Typically the process that is involved in analyzing data for the purpose of prediction entails taking a data set and splitting it into two groups: the training sample and a test sample. The data that is in the training sample is used to design the actual mathematical equation itself. The testing sample is then used to measure how good or bad the equation works for predicting new observations which did not go into the design of the original equation.

The classic example of such an analysis is the "Stock and Skirt" example. It was found in the 1970s that the average length of women's skirt sizes in fashion magazines for the year were highly correlated with the highs of the Dow Jones Industrial Average. While there is absolutely no logic as to why one would affect the other, and there most likely is no logic to support such a relationship, the theory behind the usage of the association is simple: if skirt sizes accurately predict stock market prices, then why not use it? Such analysis would not be valid if we are trying to explain stock market prices (there is no logical reason to suggest why one would affect the other!).

Yet, if our sole goal is to accurately predict stock market prices, then who cares if there is any theoretical connection between the two. Again, if it works, then it works! Therefore, how we analyze our data greatly depends on our purpose for using it in the first place. Are we using the data to describe something or to validate a theory, or are we using the data to predict

something? Depending on our intent, this will in part dictate how we will approach our data analysis. Therefore, we can begin to think about how to analyze data: namely that it begins at our goal or intent. Hence, a brief summary of purpose of data is as follows:

- Prediction (Just The "How")
  - Fitting statistical models with data such that it can predict "well".
  - Visualizing two or more variables to see what has predicted what in the past.
- Explanation (The "How" and the "Why")
  - Descriptive Analysis - characterizing a group of "things" by their characteristics.
  - Causal Analysis - fitting statistical models with data such that the model "fits" the data, and explains how the data came to be.
  - Visualizing data

Exercises

1. Briefly read the paper (no need to deeply read it, just browse it) "Effects of supply chain management practices, integration and competition capability on performance". Determine if the intent of the analysis in the paper is to explain or if it is to predict.
2. Briefly read the paper (no need to deeply read it, just browse it) "Exploring Demographic Information in Social Media for Product Recommendation". Determine if the intent of the analysis in the paper is to explain or if it is to predict.

4 Tutorial 2: What Does Data Represent?

When we gather data to either explain or predict something, the data is obviously about *something*. In order to understand how to analyze data, we first need to understand what the data itself is intended to reflect. Examples of this which are very obvious include the temperature outside or somebody's height. Other times, it is not as obvious what the data represents, or, how to represent a property with data, like trying to obtain data on "crime", "intelligence", or even "gravity". These ideas, despite us having clear images in our head of them, are not directly observable. We know that crime is "bad" and we have a general idea of what it entails, but how does one characterize or measure this with information? Likewise, we know that "intelligence" is "good", but how can we characterize this or "put a number on it"? The way in which we analyze these ideas, as well as obtain information on them, rests in first understanding on what exactly we are trying to obtain data. In this tutorial, we will briefly discuss the various concepts, ideas, and terminology that is often used for which data represents.

4.1 Units of Analysis/Individuals/Levels of Analysis

Anything for which data represents we will term as a *concept* or *entity*. In data analysis, we typically refer to a concept or entity as the "individual", which is also known as the "unit of analysis". It is the "thing" for which we seek to understand. For example, a single company may be the unit for which we seek our data to represent. A single row in our data set would represent

information about a single company. Likewise, we may be interested in understanding a single person, and hence the unit of analysis may be a single person. We may seek to understand a classroom, and hence the unit of analysis may be a classroom (i.e. a group of people, where a single row of data would represent something about a single group of individuals). Therefore, before gathering any information, or trying to understand what information in a dataset represents, the very first question we almost must ask is: "what is the unit of analysis?". Examples of units of analysis include but are not limited to:

- People
- Countries
- Groups of people
- Companies
- Networks
- Individuals in a network
- Time (we will discuss this one more in depth soon)
- States
- Cities
- Buildings
- Apartments
- Houses
- Products
- Product classes
- Departments
- Students
- Classrooms
- Schools

When the focal point of our study can be at any level of individuals that happen to be related to each other in some way, then we typically use the terminology "level of analysis", despite the meaning "individual" and "unit of analysis" being the same. For example, we could study something about education at different levels. The unit could be a single student, or we could go "one level up" to consider the focal point to not be the student, but rather a class of students. We could also go "one level up" to study not a class of students, but rather a school of students. Again, we could go "one level up" to study not a school of students, but a district of students.

The way in which we must think about our "unit" is as though if we were to gather information on a single "unit", then the information about a single unit would be all contained in a single

row in a table, where each row would be data on different units about the same type of unit. So if we were studying students, a single row in a dataset would equate to information about a single student. If our unit were classrooms, then a single row in a dataset would equate to information about a single classroom. If our unit were schools, then a single row would equate to a single school. If our unit were districts, etc...

Look at the paper "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index" under section 4. Reading the first part of that section, we see that the authors explain their unit of analysis for their project: "The research data used in this study is technical indicators and the direction of change in the daily Korea stock price index (KOSPI). The total number of samples is 2928 trading days, from January 1989 to December 1998". What is the unit of analysis in their analysis? Reading it, we can see that a single trading day for the Korea stock price index is the unit/level of analysis (aka the "individual" under study). One single row would correspond to one single observation (i.e. a single trading day of the Korea stock price index).

Likewise, we can see in Lin et al. (2018) that the unit of analysis is indicated in the research design portion of their paper: " A survey was distributed across campus to approximately 800 students (undergraduate and graduate) and staff members in a large university in the Midwest, to recruit participants from the two olfactory categories for the purpose of the study." In this instance, the unit of analysis was a single person, namely a member of the university (either a student or a staff member).

Exercises

3. Briefly read the paper (no need to deeply read it, just browse it) "Effects of supply chain management practices, integration and competition capability on performance". Determine what the unit/level of analysis (i.e. individual) is.

4. Briefly read the paper (no need to deeply read it, just browse it) "Exploring Demographic Information in Social Media for Product Recommendation". Determine what the unit/level of analysis (i.e. individual) is.

4.2 Variables, Constructs, and Measures

In any data analysis we do not only need to understand what it is we seek to better understand, but also what specifically about those individuals we seek to understand. If we think of an *entity* and how we can describe what that entity *is*, that is, to be able to distinguish it and relate it to other entities, we usually do so by looking at an entity's properties and behaviors. For example, how can I tell the difference between a group of people and a single person? Very simple: the group of people is a group, while the individual is an individual. Therefore, they are *different "things"*. How can we tell the difference between a dog and a cat? We usually use the properties and behaviors of both to distinguish between them: how each one looks, how much hair they may have or not have, the sounds that each one makes, the type of food that each one eats etc. Put simply, once we have identified *what* it is we would like to study, the next step is to identify what *about them* we would like to study.

How we characterize an individual depends on the properties about that individual we seek to better understand. A property of an individual is anything that describes that individual. Properties, however, are not always obvious to describe or gather information about. For example, let us take the human being. Some properties of a human are easy to explain and identify:

height, weight, average daily calorie intake, job occupation, number of years of education, age, ethnicity, income, etc. All of these are tangible properties, each of which are easily identifiable and measurable. They are properties for which we can *directly* observe and understand.

On the other hand, there exist properties that are not so easy to observe, measure, and understand. For example, we all have a general idea of who is "intelligent" and who is "no intelligent". Despite this, when asked to define and measure one's intelligence, people tend to struggle or they describe it in different ways. Some people will explain that intelligence is measured and described by how many things and individual can remember, while another will say it greatly depends on how applies knowledge they remembered. Definition not withstanding, the point is that a property such as intelligence is not directly observable, and we can only describe the property using properties for which we can observe. Put differently, intelligence can only be *indirectly* observed, measured, and understood.

Therefore, when we go to characterize the properties of individuals, we need to determine if the property is defined in such a manner so that it is directly or only indirectly observable. Those properties for which are directly observable are called *observables*, and sometimes depending on the context, a researcher or analyst will refer to the property as a variable. Those properties for which are indirectly observable are called *constructs*. These properties are abstract in nature, and can only be described using observables. Generally, we say that a *variable* is a property of an individual. The variable can be of the two categories just discussed:

- Types of Variables:
  - Observable: Any property of an individual for which can be directly observed, measured, and understood.
  - Construct: Any property of an individual for which cannot be directly observed, measured, and understood. It is an abstract property that can only be measured and computed in terms of observable properties.

Variables, generally speaking, go by many of the same names. You may find an analyst talking about the predictors, dimensions, properties, attributes, constructs, concepts, factors, components, or features. While some of these have specific and special meanings, this unfortunately does not stop an analyst in practice from using these terms as being interchangable with "variable". Regardless of the usage, if you happen to see any of the above words in the context of a research or analyst report, chances are they are referring to the same thing as in our discussion: it is simply a property of an individual. Here, we will consistently refer to these properties as variables, since this is the most common word which is used in practice and theory.

Taking our example from earlier, we can see that intelligence is a construct, since it is an abstract quality that only describes an individual by using other properties for which are observable. For example, intelligence can be described by the number of definitions that an individual remembers, the number of concepts or situations in which an individual were able to correctly apply the definitions. The ability to correctly deduce and induce information in a logical manner. All of these properties are directly observable, and all relate to the idea of intelligence. While "intelligence" cannot be directly observed, we can at least characterize it by other observables such as the ones above. Hence, if we have hard data on those observables, we should be able to somehow combine those numerical values together to create a *measure* of the construct, which is a number that reflects the property of the abstract concept itself.

For example, we could gather information on the number of definitions an individual remember,s the number of concepts to which they can apply the definitions, and the number of logic problems they got correct. We can subsequently use an equation to put these numbers together to get out a single score which "measures" the individual's intelligence. Higher scores may correspond to higher levels of intelligence, while lower scores may correspond to lower levels of intelligence.

In any data set, once the individual has been identified, the analyst must understand the properties they seek to study about the individual. Each property (heretofore we will refer to "property" as a "variable") must be identified as an observable or a construct. Once we know which variables under study are observable and which ones are constructs, then we know how to proceed with the *measurement* of the variables. Sometimes our data is in a *raw form*, and sometimes we have not even gathered the data yet, but we intend to. Regardless of the situation at hand, we will need to be able to assign a value to a variable for a specific individual under study.

Therefore, once we know which variables are intended to be under study and which variables are an observable or a construct, our goal is then to be able to assign a value for each variable for each specific individual. For example, if our unit of analysis is a single student, and our variable is intelligence, we then need to know how to assign a value of the property of "intelligence" for every individual student in a class. This may be as simple as using a single *Likert Scale*, whereby we ask the student on 1 to 7 scale on "intelligent" they think they are, or more advanced as carefully identifying the individual observables that help to characterize the construct and subsequently gathering and computing the measure for the construct, such as our example earlier with the number of definitions remembered and applied, as well as the number of logic problems solved.

Hence, a *measure* of a property for a specific individual is a value that we assign to the variable for a specific individual as a result of either computing the measure from other numbers for that individual, or as a result of using an instrument to gather the value. Measures have two basic properties for measure we look at to determine if it is a "good measure" or "bad measure":

- Properties of a Measure:
  - Validity - How "good" or "close" the measure actually reflects the underlying property under study. This is analogous to the concept of "accuracy".
  - Reliability - How consistent the measure is. That is, the measure's ability to report similar values upon taking and recording multiple observations. This is analogous to the concept of "precision".

Each measure can be of a few types, and need not be restricted to being only numbers. For example, one could measure the concept of income as "poor", "middle-class", "rich". One could measure the construct of firm size as "small", "medium", "large", "too big to fail". Put differently, measures can be words which hold relative comparative value. Each word would represent a category or some type of relative size. Measures of course could also be numbers, where lower and higher numbers represent lower or higher presence of the property. Hence, we have:

- Types of Measures:

  - Categorical: Each value typically represents a group within which the individual belongs. Sometimes these values have relative properties, where each word represents a general quantity or size.
  - Numerical: Each value represents a quantity or size of the underlying property.
    * Discrete: The numerical measure can be found by directly counting something.
    * Continuous: The numerical measure can be found by using a scale of some kind which can give any number from a range of numbers.

Sometimes, our variables are referred to as *dimensions* when the construct under study is a property about something that is not a person (although, sometimes the word "dimension" is used for properties of people as well). You may have heard of a group of tools which are referred to as *dimensionality reduction*. The idea behind this is very simple: how can we take measures from other variables and combine them together to get a single measure. This is the basic idea behind dimensionality reduction. Technically, there is a difference between the types of constructs and how we measure them. We have two fundamental types of constructs:

- Factor - A factor is a single measure for a construct where the observables occur as a *result of* the factor being what it is.
- Component - A component is a single measure that does not read a observable to what it is, but rather is itself a result of what the observable is.

For example, somebody's reading and math skills may high precisely because they have "high intelligence". That is, they do not have "high intelligence" because they have high math and reading score, but rather the other way around. In this instance, intelligence would be called a *factor*, since it is a construct that influences quantities for which are directly observable. On the other hand, the Dow Jones Industrial Market Index is a *component*. The Dow Jones does not cause Apple to be the value that it is, but rather the opposite. That is, the Dow Jones is itself defined and influenced by the individual stock prices themselves.

Let us look at some examples. First, let us go back to the "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index" example. We can see by reading section 4 that the authors refer to their variables as "features". Furthermore, they have a table which lists the characteristics of each variable and how it is measured. In this research, each variable is a technical indicator that is directly computed from the other information (open, high, low, close prices). All of the variables are directly observable, and their corresponding measures are continuous numerical measures.

In the paper "Understanding olfaction and emotions and the moderating role of individual differences", we can see that there are four variables: Scent, Emotion, Task, and Olfactory Group. Scent, Task, and Olfactory Group are observables, while Emotion is a construct. However, the authors used only a single measure of Emotion (namely the measurement of an electrical response in the brain as a result of emotional response). Scent, Task, and Olfactory Group are categorically measured, while emotion is a continuous numerical measure. **Exercises**

5. Briefly read the paper (no need to deeply read it, just browse it) "Effects of supply chain management practices, integration and competition capability on performance". Determine what the variables. Subsequently determine which variables are directly observable, and

which ones are constructs. Likewise, identify the measures of the variables as categorical or numerical.

6. Briefly read the paper (no need to deeply read it, just browse it) "Exploring Demographic Information in Social Media for Product Recommendation". Determine what the variables. Subsequently determine which variables are directly observable, and which ones are constructs. Likewise, identify the measures of the variables as categorical or numerical.

5 Tutorial 3: Understanding the Structure of Data

Another consideration we need to think about when we analyze data is how the data itself was generated and gathered. The concept of *time* is a very important consideration when we are trying to understand a dataset. A variable in one moment in time can influence the value of the same variable, or a different variable, at another moment in time. Therefore, we need to be careful to identify how the data on our individuals will be gathered over the dimension of time. When analyzing data, we typically have three types of data structure in terms of how the data were gathered:

- Cross-Sectional Data: The individual is defined just as we had done so above. All data for the dataset is gathered from all individuals at only a single moment in time (1 day, 1 week, 1 month, etc). The assumption behind this type of dataset is that the population will not change over the time that the data is gathered, and so any information that is inferred from the data would therefore be "accurate". Despite the fact that most datasets in practice should be gathered and account for over time, most datasets are considered "cross-sectional". The question you must ask when analyzing your data is: does the underlying population of interest change over time?

- Time-Series: Unlike cross-sectional data, where the data about a group of individuals and multiple characteristics about them is gathered only during one time frame, time-series data is a data-set that reflects a single individual and a single characteristic of that individual but over multiple points in time. For example, if you were to gather the closing price of a stock every day, the result would be a time-series, since the data is the closing price (a single characteristic) of a single company (individual). Technically, the "individuals" in the data set is a single characteristic of a single "thing" over time. That is, the "individual" is time itself.

- Panel Data/Longitudinal Data: If we were to take the idea of time-series and apply it to an entire cross section, then we obtain a new type of data set which is called a Panel Data Set. Put simply, the way in which this type of data is gathered is by taking multiple cross sections over multiple time frames. For example, we may be interested to study the open, high, low, and close prices for all stocks on the Dow Jones Industrial Average every trading data throughout the year 2019. Each individual would therefore be a single stock gathered at a specific trading day in the year. Put simply, if D represents the set of trading days and S represents the set of stocks on the Dow, then our population of individuals under study would be the Cartesian Product of Stocks over time $D \times S$.

Going back to our example papers, we notice that the authors of the paper "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index" have a data set for which is a panel data. The reason is as follows. While they are collecting information about a single individual (the Korean stock price index) over time, they are collecting multiple characteristics about that individual over time. Hence, this is not a cross section, since the concept of time is involved in the collection of data, nor is this a time-series, since more than one variable about the individual is being gathered over time. If we explore the paper "Understanding olfaction and emotions and the moderating role of individual differences", we can see that this data set is a cross section. The reason is simple: the sample were gathered in a single interval of time, and multiple observations of the same variables are not gathered. Since this data comprises of collection of multiple individuals and multiple variables, it therefore is not time-series data. Likewise, since it is not gathered over time, it is not panel data. Hence, this is cross-sectional data.

We must be careful in our analysis of data when it comes to the structure of the data and how it was gathered. If we have a straight-forward cross-sectional data set, then we need not worry about time, and we can look at interesting relationships strictly between the variables. On the other had, if we have time series data, then the only thing we can look at data wise is associations between prior data values for the individual's characteristic (i.e. variable) and current data (for example, trying to understand how the stock price of Apple on the previous trading day and the day before affects the price of it today).

Likewise, we need to be careful if our data comprises of an entire population of individuals and multiple variables gathered over time. If we were to apply simple methodologies to analyze such a dataset, we may get misleading results since the idea of time/other factors may be ignored using other methodologies. When we have such a dataset, we usually need to combine in some fashion the analysis methodologies of time-series and cross-section analysis. Such methodologies are referred to as panel/longitudinal data analysis.

Therefore, not only must we understand the goal of our analysis, the unit of analysis, the variables, the types of variables (construct vs. observable), the measures, and the measure types (categorical vs. numerical), but we must also understand the structure in which the data were gathered and represented (cross section/time-series/panel). Once we know this, then we will have a holistic understanding of how we are to approach our data analysis.

Exercises

7. Briefly read the paper (no need to deeply read it, just browse it) "Effects of supply chain management practices, integration and competition capability on performance". Determine if the data set is cross-sectional, time-series, or panel-data.

8. Briefly read the paper (no need to deeply read it, just browse it) "Exploring Demographic Information in Social Media for Product Recommendation". Determine if the data set is cross-sectional, time-series, or panel-data.

6 Tutorial 4: An Ontology of Data Analysis Tools and Data Mining Methods

Taking into account all of the prior components of a dataset, the last step is to understand what type of analysis we should conduct. Depending on our situation, and the answers to the

questions posed above, we will need to select a toolbox of methods to used from a variety of toolboxes. The toolbox we decide upon greatly depends on our purpose, variables, measures, and structure of the data we have or intend to gather. The number of analysis methods are large in number, each one intended for a specific purpose and situation within which the analyst finds themself. Despite the large number of methodologies that exist, we will attempt to present a fundamental breakdown of methods below:

- **Descriptive Analysis**
  The purpose of this type of analysis is to simply summarize the data in the data set by combining the values for a single individual variable of all individuals together. If the variables are measured numerically, then there are endless ways to combine the information into a single numerical value, since any mathematical operation of putting numbers together will suffice. There are two classes of measures that analysts commonly use to describe an entire numerical data set for a single variable: measures of central tendency and measures of dispersion. These include finding the mean, median, mode, range, min, max, variance, standard deviation, and variations of these. In addition to these two classes of metrics, analysts sometimes also compute *measures of distribution*, such as various quantiles or percentiles, and summarize these in a *relative frequency distribution table*. Visual description of the data usually entails finding histograms or box and whisker plots. If the variable is categorically measured, then the values that are usually found to summarize these types of variables are count or relative frequency tables (the number or percentage of individuals in the data set that belong to each defined category). While descriptive analysis gives us a "feel" for the composition of the entire data set, it falls short in being able to understand the association between the variables (does one variables "cause" or is "associated" or another variable?). Despite this, almost any analysis project will always begin with a descriptive breakdown of the data.

- **Basic Inferential Statistics**
  Sometimes, data collection involves the process of *sampling*. Put simply, this is the process of selecting a small group of individuals from a population, gathering data from those individuals, and analyzing their information for the intent of inferring information for all the individuals in the population (despite never observing those other individuals in the population). Such a process is referred to as *statistical inference*. Usually, inference involves finding one of two things: a confidence interval of some quantity, or the results of a hypothesis test. Statistical inference tends to be used for one of two purposes: to infer what the descriptive statistics will be, but for the entire population and not just the sample, or to infer if a relationship between two variables in the population exist. Traditional *experimental analysis* rests in this toolbox. If we managed to conduct an experiment in such a manner so that multiple groups are exactly the same, with the exception of our primary variable of interest, then we can gather the data on the two variables of interest from our sample and use the results to infer if a relationship exists. Methods to do just this include comparison of means, proportions, and general quantities between two populations, Chi-Square test of independence, ANOVA, ANCOVA, MANOVA, MANCOVA, and others. While these tests are elegant and simple to understand, unfortunately, many of them rest on unrealistic assumptions. Furthermore, too often analysts and scientists fail to check these assumptions, which can lead to very misleading results in the research project. These methods should

only be reserved for the analyst that knows they have full control over their data collection (which is almost always impossible).

- **Econometrics**
  Econometrics is a collection of analysis methods that allow the researcher to understand very complex relationships and to properly take into account situations where traditional statistical inference methods fail (which is almost, well, always). It is essentially the "king" of analysis method toolboxes. However, econometrics is a very large toolbox, and it takes the analyst quite a bit of time to master the approaches. Despite this, econometric analysis is usually best suited for when the data is secondary data, not gathered directly from the analyst, and where the goal of the researcher is to justify a theoretical model (however, the methods can equally be applied for predictive purposes). It should be noted that despite the "econ" prefix, these methods have little to do with "economics". They are general statistical approaches that have been applied not only to economic data, but also medical, political, societal, and biological areas of study.

- **Time-Series**
  Time series methods, as the name suggests, applies strictly to time-series data (data that is gathered on a single variable from a single individual over multiple points in time). The theory behind these methods is essentially that all information in the future can be predicted or ascertained from the past, and hence, there is no need to involve other individuals or variables. These analysis methods rest primarily for the purpose of prediction (although, not always!).

- **Correlation, Machine Learning, and Data Mining**
  One thing to keep in mind is that "association/correlation" and "causation" mean entirely different things. An association between two variables is when a change in one variable can be used to predict a change in another. That is, changes in variables happen to "move" with each other. Correlation, a type of association, is a *linear association*, meaning that changes in one variable will be proportional to changes in another variable. With associations, however, we cannot claim *causation* without good reason. That is, it is entirely possible that two quantities, while associated, may not be related to each other. For example, it has been shown before that the price of the stock market and the average length of women's skirts are associated. Despite this, what is the logical reason as to why? Answer: there is none. That is because there is insufficient reason and even evidence as to the *causation*. With this stated, when our goal with data analysis is purely for prediction, we usually can leverage the methods of correlation, machine learning, and data mining. This toolbox includes a variety of methods and algorithms that hold absolutely no validity for causal analysis, but do quite well for predictive analysis. Where these methods fall short is in their explanatory power. While these methods work very well for predicting quantities, they are very poor for explaining the quantities or relationships between quantities. Hence, we usually reserve our analysis to this toolbox when our goal of the use of the data is strictly for prediction.

- **Structural Equation Modeling**
  A special toolbox that deals with the involvement of constructs and observables which involve intricate and complex sets of relationships is set of methods we call *structural equation modeling*. While at the core of the analysis rests the same methods as econometrics, this

collection of methods combines some tools from *dimensionality reduction* (a subbranch of machine learning) with core econometrics. We typically reserve our use of these methods when (1) there are complex relationships between many observable and construct variables and (2) our data comes from a survey or questionnaire.

- **Textual Analytics**
  Sometimes, our individual variables will comprise of nothing by text. For example, if our individuals in our data set are "Tweets", then one variable would be "Tweet Text". This may be a very long paragraph. The question is, however do we convert the text to meaningful information? This is where textual analysis comes in. The goal of textual analytics is to extract keyword counts, common themes, topics, concepts, variables and constructs, grammar, and general feelings, emotions, and sentiments from the text. This is a toolbox that is very massive, yet still in its infancy. As the name implies, we usually conduct this analysis when our data is text, as well as if we are conducting basic elementary exploratory analysis.

- **Exploratory Analysis**
  When our goal is to explain, yet discover something in our data which we are no aware of *a priori*, then we usually conduct an *exploratory analysis*. This type of analysis involves combining various methods from the different toolboxes discussed thus far for the goal of summarizing, extracting, and discoving new information. As the same suggests, we are not trying to explain a relationship or even predict something. We are merely just trying to see "what's there" in our data.

- **Bayesian Statistics**
  Many statistical toolboxes above rest on the idea of inference. However, the methodology of inference itself rests on very questionable philosophical grounds. A competing philosophy to *frequentist statistics* (that is, the classic statistics that everybody learns in an intro to statistics course), is that of the Bayesian philosophy. Unlike frequentist statistics, which assumes that all data comes from a fixed population, and of which the goal of statistics is to use the data to say something about that population, Bayesian statistics philosophically considers data to be very different. Instead, a Bayesianist assumes that the analyst has an *a priori* knowledge about the population, and that data does not necessarily come from the population, but rather serves as *evidence* which can be used to *update* a belief about what the population actually entails. Bayesian statistics is the older form of statistics, invented in the 1800's, before the newer frequentist statistics, invented in the 1900's. However, despite its age, Bayesian Statistics is making a comeback due to its need for high computational ability. The level of accuracy for models intended for both explanation and prediction is astounding. With that said, the toolbox of Bayesian Statistics is somewhat parallel to that of frequentist, with the consistent use of *Bayes's Theorem* and *conditional probability* in statistical model specification and estimation.

Let us look at our example papers and determine which toolbox each one pulled from. For the first paper, "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", we can see that the analysis is being conducted by using an artificial neural network as well as through a genetic algorithm. These are methods that are typically common in *machine learning*. Notice how the authors do not care so much about

explaining why the technical factors of the stock impact the future price of the stock. They merely care about finding an algorithm and equation that can "predict well".

Notice the difference in our second paper, however. Recall in "Understanding olfaction and emotions and the moderating role of individual differences" that the authors were looking at a cross-sectional data set. There is also theory that is provided to justify why they believe certain relationships will hold. Observing their analysis method, we can see they undertook an ANOVA and a MANOVA. These methods belong to the traditional *Basic Inferential Statistics* toolbox. They are able to use these methods since they were conducting an experiment. Of course, as you may notice, where they have failed in their analysis is the *testing of assumptions* of the use of these methods, to determine if their data did indeed match the assumptions of the methods of ANOVA and MANOVA. Despite this, we can clearly see the authors undertook a very basic and simple statistical inference analysis.

Exercises

9. Briefly read the paper (no need to deeply read it, just browse it) "Effects of supply chain management practices, integration and competition capability on performance". Determine which of the above toolboxes are used in the analysis of this paper.

10. Briefly read the paper (no need to deeply read it, just browse it) "Exploring Demographic Information in Social Media for Product Recommendation". Determine which of the above toolboxes are used in the analysis of this paper.