Unit 1:
What is Statistics?

Neil J. Hatfield

Department of Statistics,

Pennsylvania State University

January 2020

The central question of the first unit of *Meaningful Statistics* is perhaps the deepest question of them all: What is Statistics? To start out, use the space below to quickly write out your answer to this question.

If you were to ask a dozen people this question, you'll end up with a dozen answers. However, there are some commonalities across people's responses. Figure 1 shows a word cloud for 212 people's responses to the similar question of "How would you explain Statistics?". Notice that the word "data" is the largest term, reflecting that this word was used the most often. The dots at the ends of words are for combining terms; for example, "use•" includes "use", "uses", "used", "using", and "useful". Look at your response and see which words you used appear below.

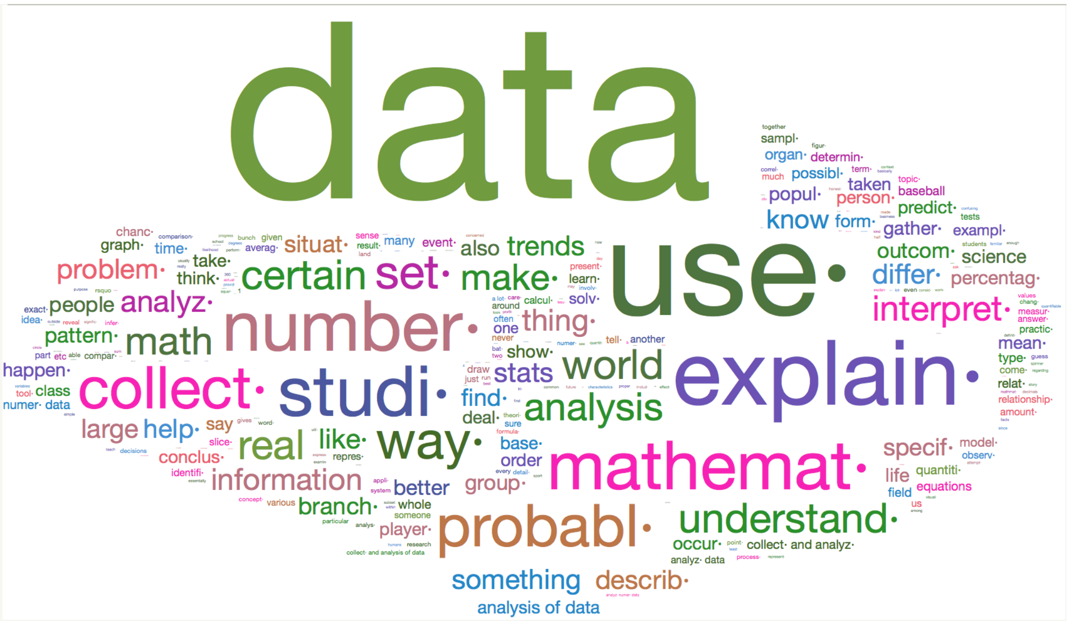


Figure 1: Word could for 212 responses to "How would you explain Statistics?"

I have seen a variety of responses to "What is Statistics?" and each have their own merits. However, there is one answer to this question that I find the most powerful:

> Statistics is the Liberal Art and Science of data to better understand our world in the face of omnipresent variation so that an individual can meaningfully participate in all levels of society.

Unpacking this definition of Statistics does take a little bit of work, but doing so is well worth our time.

Breaking down the Statistics Definition

The first thing to do when we encounter a definition is to identify the key components. Use the space below to jot down what you believe are the key components:

I believe that there are four core components: Liberal Art and Science, Data, Variation, and Participation in Society. Keep in mind that these are the components that *I* see as being essential to the definition of Statistics; you may have identified others. I must also stress that while I see these four as the core components, I do not believe that they exist in isolation and are therefore modified by the rest of the definition.

Liberal Arts and Science

The first core component, Liberal Art and Science, consists of two terms. The first term, Liberal Art, is a term that students often here at universities/colleges but rarely understand the term in any useful way. Many individuals fall into two traps when they attempt to make sense of "Liberal Arts". The first trap is that they connect "liberal" with the sense of the term in politics and the second is that they treat "art" as if they are talking about art class/time in school (i.e., the fine arts). Neither of these have much to do with the term "Liberal Arts". Rather, we have to go back in history. The term "liberal" refers to freedom and the term "art" refers to a principled practice or method. Thus, the Liberal Arts consists of the skills and ways of thinking essential for a person to be a free citizen in society. Historically, there were two core components of Liberal Arts: the Quadrivium (Music, Geometry, Arithmetic, and Astronomy) and the Trivium (Grammar, Rhetoric, and Logic). Statistics has a pride of place in the tradition of Liberal Arts as Statistics draws upon elements of both the Quadrivium and the Trivium.

Statistics is also a Science. Most people will instantly think of biology, chemistry, or physics when they think of "science" but there is much more. In essence, Science is an ongoing human endeavor to build knowledge. We create subdivisions in our endeavors to help communicate what aspects of knowledge we're trying to build. For example, biology is the endeavor to build knowledge about living organisms. Statistics is the endeavor to build our knowledge about and learning from data. Given that the Liberal Arts encompass the free, principled methods of knowledge construction, the Sciences belong here. This is why you will find Colleges of Liberal Arts and Sciences at many institutions of higher education.

Statistics is unique among the Liberal Arts and Sciences as this is a field that has a core discipline[1] and is in constant contact with nearly every other field. Hardly any other field can make such a claim. The ways of thinking that I espouse throughout this curriculum are fundamental to Statistics *and are vital to every other field.* Nearly every field makes use of data in someway: the ways in which a person reasons with and about those data are the province of Statistics. Now, I could provide examples of each Statistical concept in every field. However, none of us have that much time or space. I will do my best to provide a wide variety of examples as we go.

Data and Variation

I'll ask for you to bear with me as I'm going to skip a full discussion of both "data" and "variation" until Unit 2. The central questions in Unit 2 are "What are Data?" and "What is Variation?". For now, the intuitive meanings you have for both of these terms will not hinder you in understanding the definition of Statistics.

Participating in Society

The definition of Statistics closes with the phrase "...so that an individual can meaningfully participate in all levels of society." This phrase seems quite lofty. Given how many aspects of life Statistics touches, a productive understanding of Statistics is actually critical to being an active citizen. As a citizen of any society, you have a duty to engage in critical thinking. Statistics, being a Liberal Art and Science, provides you with critical thinking skills that will ensure that you can function as a active and informed citizen.

Back to the Definition

Now that we've torn apart the definition, we should re-examine the definition:

> Statistics is the Liberal Art and Science of data to better understand our world in the face of omnipresent variation so that an individual can meaningfully participate in all levels of society.

Take a moment to write out what you understand this definition to say:

[1] An organized formal field of study.

Personally, I understand this definition and the previous discussion of the core components to say that Statistics is a field/discipline that consists of ways of thinking that allow a person to reason coherently with varying data so that the data serves as evidence to support the person in making decisions based upon the data and not preconceived notions.

Is Statistics a Type of Mathematics or a User of Mathematics?

The tongue-in-cheek answer to this non-Yes/No question is “Yes.” In essence, the question that titles this section is asking us to decide whether Statistics is a sub-domain of Mathematics or whether Statistics is a domain more like Physics, which uses tools from Mathematics. I believe that the closest field to Statistics is Mathematics (especially with the sub-domain of Applied Mathematics). There are core components of Statistics that are firmly inline with Mathematics, but there are also aspects of which distinguish Statistics from Mathematics. Among these distinguishing aspects are the types of variation examined, the nature of problems, and the open-ended nature of questions.

In Unit 2, we’ll explore more about how Statistics deals with multiple types of variation. For now, Mathematics tends to almost always deal with a single type of variation; Statistics deals with three types of variation.

While the core of Statistics deals with problems that are similar to much of Mathematics, most problems in Statistics are grounded in a real-world inquiry (Applied Mathematics steals this approach). Often, the goal of a statistical inquiry is to build a model of a real-world phenomenon that allows us to understand a problem in a way that we can propose at least one answer to the question. David Hand once stated:

> In general, when building statistical models, we must not forget that the aim is to understand something about the real word or predict, chose an action, make a decision, summarize evidence, and so on, but always about the real world, not an abstract mathematical world. Our models are not the reality—a point well made by George Box in oft-cited remark that ‘all models are wrong, but some are useful’.

Nearly everything that we have developed in Statistics has been the result of trying to answer some question or solve some problem in the real world. As you continue through this curriculum, the real-word grounding is where the most productive meanings will often reside.

The Box quotation that Hand mentioned brings up the last distinction between Statistics and Mathematics I wish to discuss at this time. “All models are wrong, but some are useful” reflects the open-ended nature of questions in Statistics. For much of your experience in school mathematics, your experiences have led you to believe that there is one correct answer to a mathematics question[2]. Rarely is this true in Statistics. In many cases, there are multiple routes you can take to answer a research question. Each path will contain sub-paths and you'll wind up with different numerical answers and potentially different conclusions[3]. However, I must caution you at this juncture: while there is no "correct" answer in Statistics, answers must still be valid and reflect coherent, logical reasoning. Thus, whenever possible, you need to communicate your thinking.

Attitudes of Statistics

Have you ever heard the phrase "Attitude is everything" or some close variant? The sentiment behind this quotation is that the way of thinking or feeling that you bring to bear in a particular situation has a direct and dominant impact on what you're trying to do. Time and time again, we see this bear out in education. Students who believe that they aren't a "math (stat) person" end up doing worse than their peers who don't hold such a belief—never mind that there is no evidence that supports the myth of there being "math" and "non-math" people. The notion that our attitude impacts what we do also applies to how we engage in the practicing of Statistics.

Within Statistics, there are two attitudes[4] that statisticians use: Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA). The best statisticians will use both EDA and CDA as they engage in a real-world inquiry. Unfortunately, most introductory courses don't mention the existence of EDA and instead only focus on CDA. If you have ever had a Statistics class before and you spent a lot of time on Z-tests or t-tests, focusing on p-values, confidence intervals, and null hypothesis testing, you were working in the Confirmatory side of Statistics. In this course, we will discuss both attitudes but lean most heavily on the Exploratory Data Analysis side of things.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was originally developed and championed by John W. Tukey in the late 1950s, early 1960s. This attitude towards data is completely tied to the first word in the name: Exploratory. Tukey argued that at that time (and still today) there was not enough focus on just exploring what your data have to say. I must quickly point out that while data don't actually speak, a useful way of thinking to adopt is that data do speak. Exploratory data analysis demands that you set aside your personal beliefs about the phenomenon you're studying and listen to what the data have to say. This attitude is hallmarked by five core dispositions and beliefs. The central one is that ***only you*** can build your own understanding of what the data have to say. This is achieved by ***you*** digging into the data to answer the general questions of "What do you have?" and "What is going on here?".

[2] For some, your experiences have also taught you that there is only one way to get that answer.

[3] Notice that I've made a distinction between a numerical answer a conclusion; these are not the same thing.

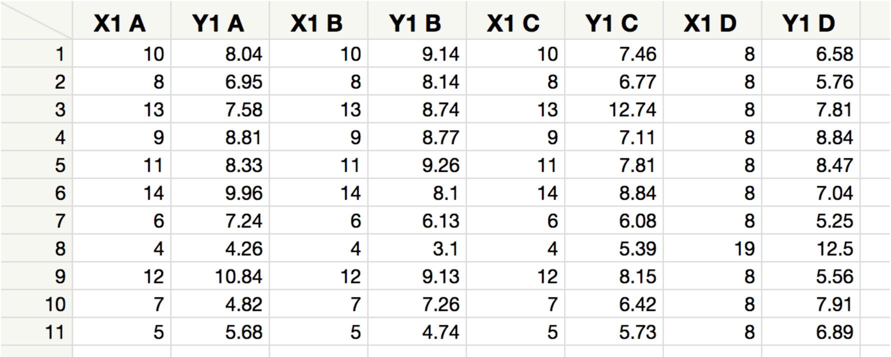
[4] Or approaches, philosophies.

Core Dispositions and Beliefs of EDA

1. The belief that you construct your understanding of the data by investigating the broad questions of “What do we have?” and “What is going on here?”
2. The belief that data visualizations play a central and vital role.
3. The belief that model building/hypothesis generation is an iterative process.
4. The belief in using methods that are robust, resistant, smooth, and have breadth.
5. A disposition of skepticism, flexibility, and statistical ecumenism for methods used.

The second most important belief is that data visualizations are central to our quest of understanding our data. As we will see throughout the course, and in particular in Unit 4, data visualizations allow us to free ourselves from our pre-existing beliefs and examine what we can learn from the data. Tukey stated that one of the greatest powers of a data visualization is that they can “force us to see what we ***never*** expected to see”.

The third belief is an immediate consequence of the first two. The third requires us to think about our construction of our understanding as a process that takes multiple steps and is ***open to revision***. No one can develop “the perfect” understanding of any particular data set or phenomenon in a single moment. Rather, we build our understanding a piece at a time. As we try to understand some phenomenon/problem in the real world, we use data. However, we must acknowledge that we could have a completely different data set and be open to revising our understanding. While this openness to revision might strike some people as indicating that Statistics isn’t an “exact science”, this openness is what makes Statistics a science. As we explore data, we might form one particular understanding, but as we use a different tool, our understanding might change. To exemplify this, examine that data table below (Figure 2).



	X1 A	Y1 A	X1 B	Y1 B	X1 C	Y1 C	X1 D	Y1 D
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.81
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Figure 2: Data Table for Anscombe’s Quartet

We can think about each column labeled "X1" refers to the same underlying attribute upon which we took measurements; similarly for each column containing "Y1". The A, B, C, and D mark out four separate groups. Take a moment and jot down some of the things that you notice about the Quartet.

Since we're in Statistics, most people will immediately reach for tools they associate with Statistics: typically, the "average" (i.e., the *sample arithmetic mean*). Those with a bit more experience will also reach for more sophisticated tools such as simple linear regression. In terms of the X1 values, all four groups have the same value for the *sample arithmetic mean* (9); the same is true for the Y1 values, each group's value of the *sample arithmetic mean* is 7.5. For those who used simple linear regression, all four groups produce a regression line of the form $Y = 3 + 0.5X$ with a correlation coefficient of 0.82 and an $r^2 = 0.67$. Thus, we might think that while we have four groups, the same kind of relationship occurs in each. If we have an EDA attitude, we would not be satisfied with this. Rather, we would want to dig into the data deeper. In fact, if we were to plot the data, our understanding of the quartet drastically changes. In fact, Figure 3 provides us such an opportunity.

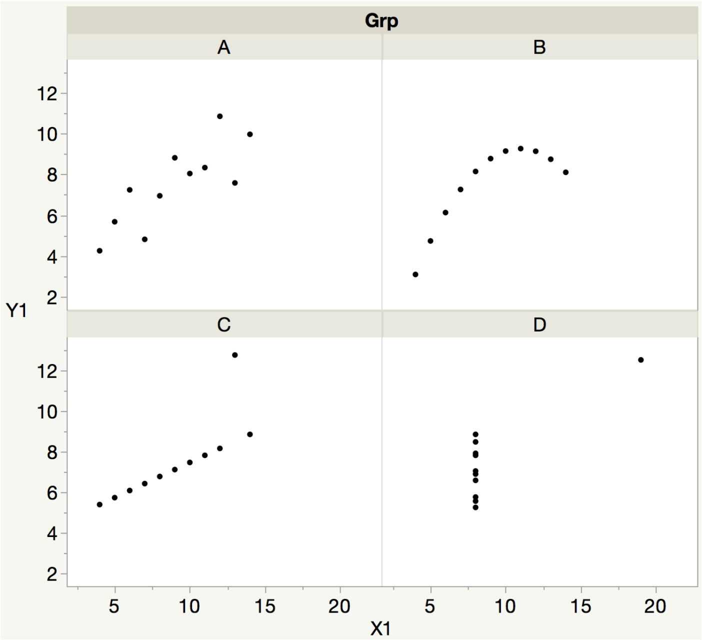


Figure 3: Side-by-Side Plots for Anscombe's Quartet

We can now see that each of the four groups have distinct behaviors to the underlying relationship between X1 and Y1. By incorporating the first two beliefs, the third allows us to build models of situations that reflect our ever developing understanding of those situations.

Given that the EDA attitude keeps us open to revision, we need to adopt a stance towards using tools and methods that won't steer us in problematic directions. This is where the fourth core belief comes into play. Nearly every tool we use in Statistics hinges upon assumptions that we make, even tools such as histograms, boxplots, the *sample median*, and *sample arithmetic mean*. When the data and our thinking about the data don't support these assumptions (i.e., the data/we violate the assumptions), then the results of the tools are questionable. The **robustness** of a tool/method/statistic refers to that tool's ability to handle us breaking the underlying assumptions and still return something that makes sense.

When you have a data set, you might find yourself looking that the values of individuals who don't seem like the others. For instance, the dot in the upper right corner for Group D of Figure 3, or looking at the heights of students in a class that also contains Shaquille O'Neal. These individuals are often referred to as outliers[5]. Another case that you might have is that your data values are always subject to measurement and observation errors. Data, whether collected by a machine or by a human, are subject to errors. In both the case of outliers/outsiders and errors, the result of using our tools are subject to influences from these values. A tool's ability to mitigate the potentially undue influence from these values is what we refer to as the tool's **resistance**. The more resistant a method is to outliers/errors, the more we can trust our understandings based upon these methods.

Similar to the notion of resistance is the idea of smoothness. The **smoothness** of a tool is the ability of the tool to handle the introduction of "bad data" (i.e., values we know aren't valid for the situation). When we develop a new tool, we use simulations to test the tool's smoothness. This allows us to build up our confidence that the tool can handle a data value that does not make sense but we haven't noticed yet.

The last aspect, **breadth**, refers to the number of different kinds of situations a particular tool/method/statistic can be used. Most of the time, we developed a tool to solve a specific problem. Then, we test out that tool in new situations; if the tool works, the breadth has increased. The *Count* statistic has the widest breadth of all as this statistic can be used in *any* situation. The *sample arithmetic mean* on the other hand gets used in a lot of situations where the result makes no sense[6].

Essential to the attitude of EDA are the dispositions of skepticism, flexibility, and statistical ecumenism. Just so that everyone is on the same page, a disposition refers to a routine aspect of a person's thinking or character. When we say that someone has a "sunny disposition", we mean that that person is often cheerful and has a positive outlook on life, regardless of what might be going on. When you embrace EDA, you maintain a degree of both skepticism and flexibility to your thinking. The skepticism will ensure that you adhere to both the first and third core beliefs. The flexibility will help you embrace the first four beliefs. The last disposition, statisti-

[5] There is some debate about whether they should be referred to as "outliers" or as "outsiders", given we can never truly know whether someone is an outlier.

[6] I once saw a group present data and reported the value of the *SAM* for biological sex.

cal ecumenism, is a difficult one for many individuals who grew up in the American school system. This disposition requires that you don't view mathematics as a tool that tells us the correct and complete answer. Rather, mathematics is a tool that we can use to solve a problem, and we judge that tool on how helpful that tool is to us as we build our understanding of the data.

Taken together, these five core beliefs and dispositions allow a person to fully adopt the EDA attitude and gain a high degree of mastery in Statistics.

Confirmatory Data Analysis

The second attitude in Statistics is that of Confirmatory Data Analysis (CDA). Just as "Exploratory" holds the soul of EDA, "Confirmatory" does the same for CDA. This is that attitude that we need to adopt when we want to test out whether a particular model we have for a phenomenon works[7]. The notions of null hypothesis testing and interval construction are the provenience of CDA[8].

Even though Tukey championed that EDA be explicitly taught starting in the 1960s, and several others have echoed his call (e.g., Behrens, and host of Statistics Education researchers), few curricula have moved towards the adoption of EDA. Almost all Introductory Statistics curricula are stuck squarely in the CDA attitude.

Using both EDA and CDA

To have true mastery in Statistics, we must have and use both attitudes. To only use one or the other is to limit ourselves. Using CDA in the absence of EDA is to go against the history of scientific method and leads us to the problems of the current replication crisis. However, to use EDA without ever moving to CDA deprives us of making decisions about real-world situations. Remember, Statistics is grounded in addressing something that is happening; EDA helps us build models but stops short of saying what we need to do. CDA on the other hand, allows us to test out the EDA models so that we can reach a decision for what to do.

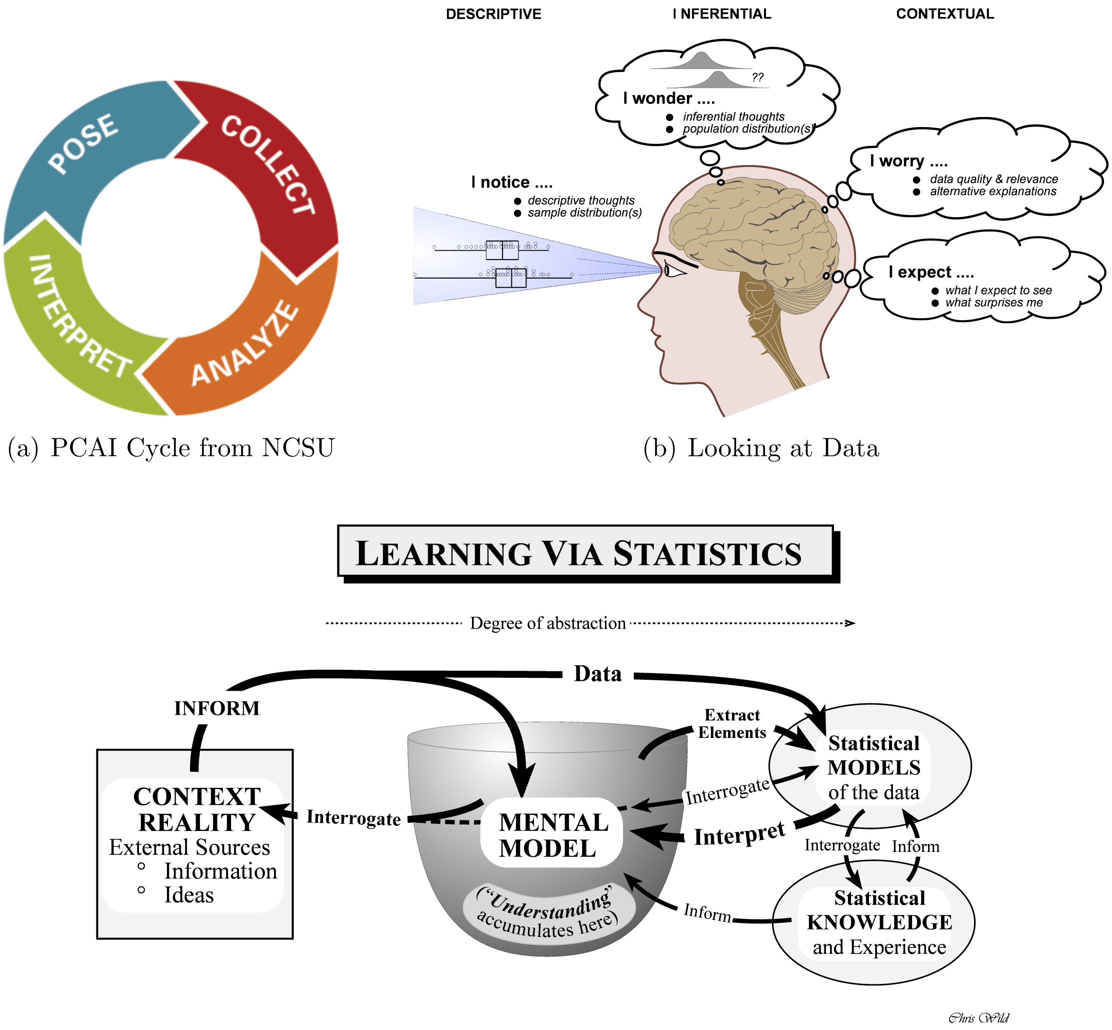
A useful way to think about the relationship between EDA and CDA is through the analogy of the TV show *Law & Order*. EDA reflects the work of the detectives as they try to piece together a model of the episode's crime. CDA reflects the jury trial component of the episode. However, there is a critical distinction between the analogy and the actual use of EDA and CDA. In real life, the evidence found/used by the detectives and the prosecutor is the same; in Statistics, we must use two different data sets. The first data set helps us build our model and the second data set allows us to test that model. The two most popular ways to achieve this method are 1) collect data in two rounds, and 2) to collect one set of data but, subdivide the data into two pieces (one for EDA and the other when you finally move to CDA).

[7] Keep in mind the "works" and "is True" are NOT the same thing.

[8] However, these tools can be used with the EDA attitude.

The Statistics Workflow

If you are to engage in the practice of Statistics, then you must begin with recognizing the flow of doing Statistics. There are myriad of different ways that people have sought to describe the practice of Statistics in visual form. Three such approaches (Figure 4) include the PCAI Cycle[9], Looking at Data[10], and the Learning via Statistics[11]:



(a) PCAI Cycle from NCSU

(b) Looking at Data

(c) Learning via Statistics

Figure 4: Three examples for visualizing doing Statistics

Each of these three cycles has pros and cons. For example, the PCAI cycle has the pro of being the simplest to look at; Looking at Data places an emphasis on data visualizations; Learning via Statistics grounds everything in reality. Both the PCAI

[9] From North Carolina State University, Teaching Statistics Through Data Investigations

[10] Pfannkuch, M., Regan, M., Wild, C. and Horton, N.J. (2010) Telling Data Stories: Essential Dialogues for Comparative Reasoning. Journal of Statistics Education, 18(1).

[11] Wild, C.J. and Pfannkuch, M. (1999) Statistical thinking in empirical enquiry (with discussion). International Statistical Review, 67, 221-266.

and Learning via Statistics highlight the cyclical nature of doing Statistics. However, all three hide the interplay and role of the two attitudes of Statistics; that is EDA and CDA. Figure 5 shows my interpretation of how a person engages in the practice of Statistics.

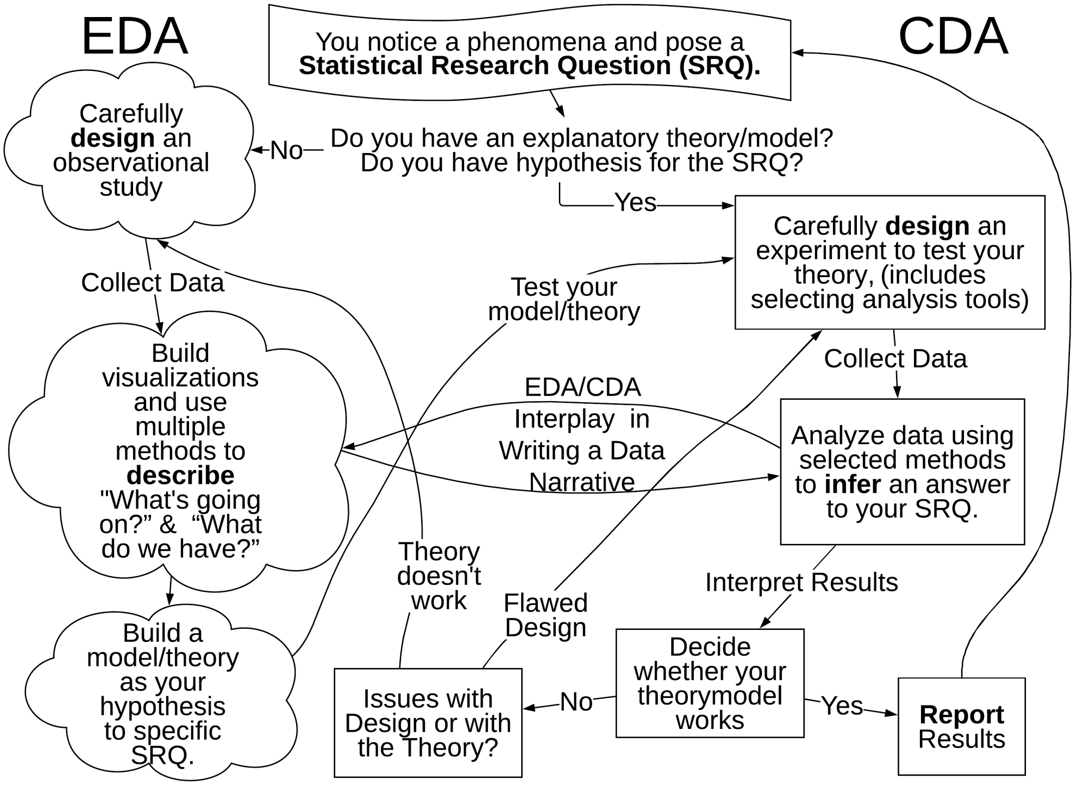

EDA
CDA
You notice a phenomena and pose a
Statistical Research Question (SRQ).
Do you have an explanatory theory/model?
Do you have hypothesis for the SRQ?
No
Yes
Carefully design an observational study
Carefully design an experiment to test your theory, (includes selecting analysis tools)
Collect Data
Test your model/theory
Collect Data
Build visualizations and use multiple methods to describe "What's going on?" & "What do we have?"
EDA/CDA Interplay in Writing a Data Narrative
Analyze data using selected methods to infer an answer to your SRQ.
Theory doesn't work
Flawed Design
Interpret Results
Build a model/theory as your hypothesis to specific SRQ.
Issues with Design or with the Theory?
No
Decide whether your theorymodel works
Yes
Report Results

Figure 5: The Statistics Workflow

The cloud shaped boxes reflect the where the EDA attitude needs to be the dominant, while the rectangles reflect where CDA is dominant. As the Statistical Research question (SRQ) encompasses both EDA and CDA, this appears as a wavy rectangle. This workflow highlights that the source of inquiry is from your noticing of some phenomenon (i.e., a real-world grounding) as well as how EDA and CDA play into each other in a dynamic way. There are five key phases to the workflow[12]: the Statistical Research Question, the Study/Experiment Design, Data Description, Statistical Inference, and Communication/Reporting and Reflection[13]. Each one of these phases help us to answer our Statistical Research question, even if the answer is another question. While this workflow may look complex or messy, this is a reflection of the nature of Statistics and problem solving.

[12] Listed in bold typeface in the figure.

[13] Other names for these phases are Problem, Plan, Data, Analysis, and Conclusion.