Revisiting statistical analysis curriculum in a data era: a learning-by-mistake approach

Gabriela Duque, Ana; Gonçalves Melo Pequito, S.D.; Rosado Coelho, Joana

Publication date
2020

Document Version
Final published version

Revisiting statistical analysis curriculum in a data era: a learning-by-mistake approach

Ana G. Duque Schumacher, Sérgio Pequito, Joana Rosado Coelho

***Abstract*— Contribution: We re-think the 'Statistical Analysis' curriculum building upon system engineering tools where assumptions (e.g., ABET criteria and student profiles) are carefully assessed, a learn-by-mistake approach ensures that several of the main statistical mistakes are learned, and advanced topics are proposed to make a strong connection with forthcoming courses in data science.**

Background: Today's data science requires students and prospective data scientists to have a strong foundational background in statistical analysis methods and decision making. Given the diversity of students' profiles, and a multitude of statistical analysis curricula across the USA, we seek to provide guidelines on a curriculum that is in line with today's data demanding era.

Intended outcomes: The target audience comprises students in engineering courses that deal with data and seek to obtain a domain-specific technological or societal solution. Using a learn-by-mistake approach, we try to mend some of the most common mistakes in statistical analysis for the new generations of data professionals. The proposed curriculum equips students with multiple statistical methodologies that enable them to understand, process, extract, visualize, and communicate statistical evidence.

Application design: We propose a systems engineering approach to design the curriculum that leverages tools and methodologies from operations research and statistics.

Findings: Our approach ensures that the designed 'Statistical Analysis' course satisfies some of the intended constraints and goals by design. In particular, we designed an overarching hands-on example that integrates the topics covered in the curriculum into a transversal example and can be further customized to the different students' profiles.

***Index Terms*— Curriculum development, Statistical Analysis, Data science, Undergraduate education, Industrial engineering**

I. Introduction

In this manuscript, we provide a study and a reflection on the curriculum design of a course often termed "Statistical Analysis" which is an undergraduate introductory course to data science – an inter-disciplinary field that uses statistical methods, processes, algorithms, and systems to extract knowledge and insights from many structural and unstructured data. Statistical analysis encompasses the study of the organization and interpretation of data following an established experimental method or procedure [1]. Therefore, the course objective is to equip the students with statistical methodologies that allow them to understand, process, extract, visualize, and communicate statistical evidence obtained. Ergo, the course is a stepping stone for other data science courses as well as social science or engineering courses that deal with data and seek to obtain a domain-specific technological or societal solution that lies at the core of systems engineering.

A. Goal

Our study seeks to revisit the curriculum of the Statistical Analysis courses (e.g., the 'ISYE-4140: Statistical Analysis' taught at Rensselaer Polytechnic Institute (RPI)). This course is a required course for most engineering courses and it pre-requisites an initial course in probability and statistics (e.g., "ENGR-2600: Modeling and Analysis of Uncertainty" taught at RPI), where the topics covered include Descriptive Statistics, Probability, Discrete Random Variables and their Probability Distributions, Continuous Random Variable and their Probability Distribution, Joint Probably Distributions and Random Samples, Point Estimation, Statistical Intervals Based On Single Sample, Test of Hypothesis Based On Single Sample, Inferences Based On Two Samples. Additionally, an overview of advanced topics is likely provided that may include ANOVA, Simple Linear Regression, and Goodness-of-Fit tests.

B. Methodology

In our study, we propose to employ a systems engineering perspective on the course design that leverages tools and methodologies from operations research and statistics. First, at the high-level, we seek to create a curriculum structure that is in-line with today's data-demanding era. Additionally, we propose a *learning-by-mistake* approach that seeks to *avoid some of the most common statistical mistakes find in the literature* [10]. The most common mistakes made when it comes to statistical analysis provide us with an insight into the knowledge gaps that exist, which we try to mend for the new generations of data professionals.

In particular, we believe that this course will become a compulsory introductory course in data science in any social and engineering student's curriculum. As such, it will equip them with invaluable skills for today's and tomorrow's job market. Second, we will pose *assumptions*, *constraints*, and *goals* to be achieved that will serve as a proxy to a problem statement that implicitly attains the aforementioned high-level

The authors were previously with the Industrial and Systems Engineering Department at Rensselaer Polytechnic Institute, Troy, NY 12180 USA. A. G. Duque Schumacher (anagabrieladuques@gmail.com). Sérgio Pequito (sergio.pequito@tudelft.nl). Joana Rosado Coelho (jrosadocoelho@gmail.com). Sergio Pequito currently with the Center for Systems and Control, Delft University of Technology, Netherlands.

goal. To some extent, our proposed approach implements in a systematic manner some aspects of constructive alignment [16], [17]. Specifically, we proceed backward (in a control theory-like fashion) from Learning Objectives (i.e., what they should learn) to Learning Activities (how they should learn). Yet, we do not discuss in detail the Learning Assessment besides the fact they should replicate the knowledge and exercises produced during the Learning Assessment. Furthermore, we notice that we seek to attain Learning Objectives in a direct, as well as by contradiction, manner. In other words, by showing what we can obtain by using given statistical tools, but also by emphasizing that what we obtain has limitations that need to be circumvented. Consequently, by introducing a *learning-by-mistake approach* in the Learning Activities, we believe that our approach will ensure that the re-designed 'Statistical Analysis' course satisfies some of the intended constraints and goals by design.

C. Assumptions

The central assumption is that the topics in the introductory class in probability and statistics are fully covered. As an implicit constraint, the course must achieve the ABET Student and Learning Outcomes, using the criteria for accrediting engineering programs effective for review during the 2019-2020 accreditation cycle [2]. An explicit constraint we took into consideration was student population profile (i.e., what job they will perform and in what capacity will data science be present in their daily affairs) described hereafter in the context of industrial and management engineering (IME).

Regarding the goals, the students should satisfy the seven ABET student outcomes that establish the abilities that should be acquired through coursework to fulfill an engineering degree [2]. Another goal is that students across different students' profiles should be able to pose relevant questions that can be statistically tested, for which they have to gather and organize the required data, and ultimately, translate the results obtained to a broader audience. Lastly, this course should enhance students' awareness of limitations and some major methodological flaws, which will increase the potential of good practices to be used in industrial, academic, and research professions.

Another assumption is that statistical analysis is a course that meets twice a week for one hour and fifty minutes for 14 weeks that constitutes the semester, a total of 28 lectures. Additionally, we considered and compared similar courses from multiple universities and colleges across the United States in a systematic manner to establish a well-rounded course structure. Besides the common threads present in these courses, we found unique topics that are (in the authors' opinion) spearheading the most innovative data science courses. As such, we seek to ensure that the designed course is teaching skills and techniques needed in the industry, rather than solely the classic topics taught traditionally in these types of classes.

II. Student Profiles

As statistical analysis is (in particular) taught to IME students, it is imperative for the course to be catered to the job market needs, while building a strong foundation in statistical analysis. Industrial engineers hold a little over 280,000 jobs in the United States as of 2018 [3]. The largest employers of industrial engineers are transportation equipment manufacturing, followed by computer and electronic product manufacturing and then professional, scientific, and technical services [3]. With a predicted growth of 8% in the next ten years, the need for industrial engineers is growing. It is important to note that the market for data scientists is expected to double in the next five years [4]. Thus, highlighting the importance of students to have a strong set of data science (and, in particular, statistical analysis) skills which will make them distinguished assets in the marketplace.

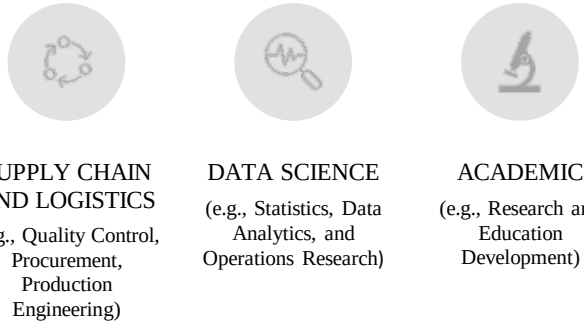

Fig. 1. Student Profiles.

To be able to cater to those needs, we categorized the most common pathways taken by industrial engineers nationwide as well as for students at RPI. The three main overarching profiles created are as follows: *(i)* supply chain and logistics; *(ii)* data science; and *(iii)* academic - see Fig. 1. Notice that the profiles highlighted in Fig. 1 overlap in many industries, skills, or topics, but differentiate in the overall pathway taken by professionals. Therefore, statistical analysis can be considered as the last stop before students branch into their specific interests. As such, not only must it cater to these profiles, but also ensure it properly prepares the students for future courses that further deepen concepts taught in the class.

In fact, the students' profiles also represent the students' broad areas of interest. Therefore, the course curriculum and the class setup should leverage these interests to capture and maintain the students' attention and engagement. An alternative is to provide multiple examples that represent real-life situations that might be faced by each type of profile. As such, by creating a curriculum that tries to mimic real-life situations faced by industrial engineers, the material will be more applicable and more relatable. Research suggests that case-based examples are useful in facilitating student understanding [5], [6]. In fact, there has been a prevalence of these in statistics courses [7]. This practice has been implemented in a multitude of courses due to the lack of a direct link to industrial problems, which constitutes an obstacle for the students that ultimately leads them to fail to comprehend the real-life applications of statistical concepts [8].

III. Course Setup proposal

We propose the course to be a hybrid between (traditional or virtual) classroom lectures and (hands-on) laboratories. In this setup, lectures will resort to slides presentation of the concepts and examples, and intertwined with labs where the software suggested is the programing language R. R is widely used in data analysis. It is a free open source software package that allows for multiple packages to be easily downloaded to cater to the different topics covered in class and the marketplace. It is compatible with many commonly used platforms. Furthermore, it is often requested as a programming language in a variety of job posts in the different students' profiles described. Although it is one of the most popular tools in analytics, the course is designed to ensure that students can be flexible and apply the knowledge taught to other data software – as mentioned by different directors from four of the nation's top university analytics programs, it is of utmost importance that students are able to learn a new tool over the importance of mastering one specific software, as the software selection is continuously evolving over time [9].

The hybrid setup will give the opportunity to teach the conceptual and theoretical content while allowing for examples and demonstrations to be done interchangeably with the lectures. The labs will be used to work out a variety of examples that will provide experience and insights on the concepts introduced during the lectures. The intention is for the course to switch between these two outlets when it is most applicable and prevent harsh disruptions in the content flow. The examples worked out during the labs should reflect the different students' profiles such that students relate to what is being taught in the lecture and keep engaged.

IV. Course curriculum and structure

As previously mentioned, we will adopt a systems engineering perspective to the problem of deriving a curriculum for Statistical Analysis that satisfies all the constraints previously mentioned. Towards finding a solution to the problem of re-thinking the Statistical Analysis curriculum, we first seek to understand the potentially feasible solutions. Therefore, we compiled a list of the topics covered in similar courses that are preceded by an introductory class in probability and statistics. This list of topics suggests that some core topics are essential. Specifically, bootstrapping, ANOVA, Simple Linear Regression, and Multiple Linear Regression. Furthermore, this list also revealed emerging topics that are being introduced in the curriculum as the industry and the field of data science is evolving (e.g., Non-Parametric Testing, Categorical Data, and *p*-curve).

Secondly, we perturbed a potentially feasible solution to accommodate some of the additional constraints in our problem formulation (i.e., the need for the awareness of limitations and misguided conclusions due to mistakes in using statistical tools). The main reason for this constraint is that software tools widely used in decision making often fail to provide 'informed consent' on the assumptions required to draw a conclusion, and the limitations are often hidden which can appear as the statistical language that is 'too technical'; thus, it is not uncommon to see it replaced by vague and abusive terminology that compromise the inferences and conclusions obtained in facto.

As such, we propose a *learning-by-mistake* approach that seeks to *avoid some of the most common statistical mistakes find in the literature* [10]. The most common mistakes made when it comes to statistical analysis provide us with an insight into the knowledge gaps that exist, which we try to mend for the new generations of data professionals.

Again, we believe that by bringing awareness to the most common mistakes and how to prevent them, the course will be able to mitigate what is currently occurring in the field of data science and, in particular, several disciplines in science and engineering. The leading mistakes are transversal and occur due to the following: *(i)* lack of proper experimental setup; *(ii)* absence of adequate control groups; *(iii)* small sample sizes; and *(iv)* the tendency for circular analysis (i.e., distorting the results of a statistical test by retrospectively selecting features within the data to characterize the dependent variables).

With these mistakes in mind, the class can focus on emphasizing these topics when teaching all topics covered. Simply speaking, these mistakes are tied with the following: *(i)* what are the hypotheses being formulated (i.e., the null and the alternative hypotheses) and how we are going to test them; *(ii)* statistics are used to refute the hypothesis, so if we are not able to refute the hypothesis we posed, it does not mean that we should accept it, yet we can try to have a control group where this hypothesis will be rejected which will not make our initial hypothesis true, but it would strengthen the evidence that is more likely given the data considered; *(iii)* as each conclusion of a statistical test depends on a specific data collected, the conclusions become more sensitive to smaller data samples considered; and *(iv)* we cannot use data to try several hypotheses that were not initially considered as the first two points might be compromised (among others). The next group of common mistakes is associated with hypothesis testing and related to inflation in the units of analysis and spurious correlation. These mistakes will be revisited when overviewing hypothesis tests as well as when these are used in the context of correlations and ANOVA – see [15] for the detailed course schedule.

With the above structure in mind, the first class is used to open the course by introducing the professor, the overarching purpose of the class, overviewing the syllabus, and evaluation criteria. When presenting the overarching purpose of the class, the students' profiles designed should be considered to provide multiple examples associated with situations seen in the career paths taken by each profile, see Section V-D for examples.

After the introductory class, the next two classes are designed as a review of concepts from the first introductory course in probability and statistics. Specifically, review on graphs and the importance of graphing data. Here, we should emphasize one of the most common mistakes identified (i.e., the spurious correlation) by displaying the Anscombe's quartet that emphasizes the importance of displaying data and the correlation assumptions. Additionally, topics reviewed in these

two classes should entail discrete random variables and probability distributions, continuous random variables and probability distributions, joint probability distributions and random samples, statistical intervals single sample, the test of hypothesis with a single sample, and inferences based on two samples. One of the highlighted mistakes that should be reiterated here is interpreting comparisons between two effects without directly comparing them. Here *t*-test (unpaired *t*-test) can be used as an example, specifically, by having a case in the R lab where the mistake is made and then propose the right methodology to correct it – the proper way of comparing the two effects. It is recommended for the Teaching Assistant (TA) of the class to hold that week a more in-depth optional review, which will enable students that do not have such a strong background on the topics to have the opportunity to catch up at the beginning of the course.

After the review, we suggest introducing the concept of bootstrapping as it circumvents one of the major mistakes when dealing with a small number of samples and enables us to detect spurious correlations. Good references for the introduction of this topic are reference [11], [12]. Next, we propose the introduction and full elaboration of ANOVA as it builds upon the hypothesis tests reviewed in the first classes. In particular, we will emphasize the role of multiple hypothesis testing and the fact that ANOVA seeks to establish if there is statistical evidence that all objects of study have a similar property versus if there is at least one that does not have it. Simply speaking, a common logic flaw is to think that the negation of all is none, but it is instead that there is at least one that does not satisfy a property. Hence, if there is at least one that does not satisfy the property, then multiple pairwise tests must be conducted to determine which hypothesis can be rejected.

This would be a good opportunity to consider introducing the lack of proper experimental setup, as well as highlighting mistakes like the absence of an adequate control condition group, circular analysis, failing to correct for multiple comparisons, and the use of small sample size. With the mistake of the absence of an adequate control condition/group, it is recommended to expand and give an example of the impact that bias has on experiments. In line with this, it would be good to introduce double-blind experiments and their limitations. A circular analysis example is also recommended to provide guidance on the importance of establishing a hypothesis before the experiment setup while expanding on the concept of double-dipping. By referring to the common mistake of multiple comparisons, we can introduce the concept of Family Wise Error Rate, and the effect of exploratory research, mentioning Bonferroni correction and false discovery rate. Lastly, we recommend assessing the impact of failing to correct for multiple comparisons, as well as reviewing the effect of sample size and its relationship with Type II error through simulations in R labs.

Following ANOVA that seeks to establish comparisons, we recommended transitioning to simple and multiple linear regression that aim to establish dependencies. In these topics, we suggest highlighting mistakes like inflating unit of analysis, misinterpreting results on correlation, and erroneously replacing correlation with causation. For inflating unit of analysis, the proper selection of degrees of freedom and the expansion on how results change based on the degrees of freedom used can be shown in an R lab example. Additionally, we seek to introduce the Simpson's paradox and several examples of it. Simpson's paradox is a phenomenon observed when correlation in several different groups of data disappears or reverses when these groups are combined. Lastly, correlation and causation should also be highlighted to bring awareness that correlation does not imply causation and the effects and repercussions of wrongfully considering doing so [13].

The last topics of the course are non-parametric testing, bootstrapping methods, categorical data, and *p*-curve. All of these are considered advanced topics that would be the stepping stone for expansion in future classes. Notwithstanding, an interesting alternative would be to consider how the different topics would adapt to a situation where categorical data is also deemed – as this leads to some of the most common mistakes found in practice (e.g., the Simpson's paradox). We would like to take the opportunity to emphasize the importance of a less-known concept of *p*-curve that provides an opportunity to expand on common mistakes like flexibility analysis and over-interpreting non-significant results. Specifically, it measures the variation of the *p*-value obtained as a function of the sample sizes considered which dependency should follow a power-law-like distribution. Thus, precluding brute-forced data crunching to attain a relevant *p*-value used to claim statistical significance – also known as *p*-hacking. As such, an R lab could consist of providing data that enables one to do *p*-hacking, and later use the *p*-curve to diagnose its existence.

V. Hand on data example

We have designed an overarching hands-on example to illustrate how the different concepts can be integrated into a transversal example and adapted for the different students' profiles. The example is based on a variable "X" – this example was created as a format for customization pertaining to a desired topic in line with the students' profiles. In our case, the example stems from reference [14] with simulated data, where "X" is a plant being grown in the United States; specifically, industrial hemp that has just recently been introduced into the US market as an emerging and innovative plant. This example navigates through the topics taught throughout the semester in a general sense and connects the dots with one overarching example. Briefly, the first part (Part I) of the problem guides students through visualization and conducting *t*-tests between the relevant variables of the data set. The second part (Part II) of the example presents more data that is collected in relation to costs associated with growing the plant. Initially, grouped data is provided to find insights in relation to the regions using ANOVA. The third part (Part III) of the example provides a digression over four different settings and linear models that allow for students to see the applications and limitations of linear models using different weights descriptions and error characteristics.

In what follows, we provide a detailed description of how the data was created within different settings, whereas the description of the statistical methodology being deployed, and

the conclusions drawn are relegated to the supplementary materials. This decision is made to emphasize the reasoning behind the proposed exercises, while enabling the reader to build upon the different examples to be adapted to the different students' profiles. The overarching example navigates through various topics covered in the course, segmented into three parts. [15] provide R Markdown files that walk through all the problem description and development. The simulated data files that were used specifically to develop this problem can be viewed in the R Markdown provided in for each of the parts of the problem in the corresponding supplementary materials. Each data file has a specific name (e.g., Price_Data_2019_2030). Further details on how the data was generated are described in this manuscript.

Consider the following description to establish the bigger picture of the example: *In 2018, research was conducted to find the price of "X" seeds per metric ton. "X" is a plant that would be introduced in the US market in 2019. The research compiled information from all over the United States and gave a low, medium, and high price for metric ton of "X" seeds. The research was conducted to gather data concerning the cost related to growing this plant. Note: One should consider a significance level of 0.05.*

A. Part I

The first part of the example pertains to the importance of visualization and the need for hypothesis tests. In 2019 when "X" was introduced to the US market, the selling price was recorded for each state. In 2030 when "X" has been part of the economy for more than 10 years the selling price was recorded for each state. We want to compare these with the costs predicted found in research conducted before 2019. It established a high price for metric ton of "X" seeds of 19,841.62, medium price metric ton of "X" seeds of 8,818.50, and lastly, a low price metric ton of "X" seeds of 4,406.06. Data (Price_Data_2019_2030) was simulated for both 2019 and 2030 with one data point per state using a normal distribution with a mean of 6000 and a variance of 2000 for 2019, and for 2030 a mean of 4400 and a variance of 100 – see Table 1. The three expected data points are extracted from reference [14]. All the results are presented in reference [15].

TABLE 1
DATA FOR SIMULATION OF PRICE 2019 AND 2030

Year	Mean	Variance
2019	6000	2000
2030	4400	100

The first part of the example guides the student into visualizing the data and highlights its importance. After gathering some insights on the data, the first step in the example involves looking at the data collected for 2019 and analyzing whether either the expected low, medium, or high data differ significantly from the data provided in 2019. Therefore, a *t*-test is conducted with each of the expected values for 2019 and 2030. As can be seen in the detailed example in reference [15], all of them are statistically significant to the data expected for 2019, with a significance level of 0.05. Furthermore, we can assess if, in 2030, the data collected allows for a different conclusion (i.e., would it make clear that some hypothesis of an expected value is likely to be rejected). When conducting the *t*-tests, the student can see that in the case of the expected low value, the data is not statistically significant for 2030, with a significance level of 0.05. Notice that this is not surprising as it is in-line with the intuition, given that the 2030 data collected has lower variance. As mentioned previously, the significance level for this whole problem set is 0.05. It is important to highlight the impact that different significance levels have on the conclusions in an experiment. A relevant exercise can be done in class on providing different significance levels and taking the corresponding conclusions. These exercises allow the students to be able to see that no matter what data is used, there is always a significance level for which a claim will be true, shining light on the importance of establishing the significance level properly before the experiment is conducted. Although different conclusions will be developed with different data, the same steps can be taken to drive the students to understand the overall example.

B. Part II

The second part of the example would state: *When the "X" seed was introduced into the market in 2019, it was purchased and grown by many farmers. One farmer per state reported the three main costs associated with growing the plant: A, B, and C. The states were separated in regions 1-5. All these costs were summed into the total cost of growing the plant.* All the results can be seen in reference [15].

This example will guide the student towards using ANOVA and understanding its limitations. Starting with grouped data (Regions_Data), we created 5 regions out of the 50 states (i.e., 10 states per region) – see Table 2. In this case, data was generated for each cost. In the case of cost A (corresponding to the cost of seeding in our example), the costs for different states were generated as follows: for region 1, 2 and 3 with a normal distribution with a mean of 1300 and a variance of 500, whereas for region 4 and 5 we considered a normal distribution with a mean of 600 and a variance of 200. Additionally, the cost B for all different states (corresponding to the cost of fertilizer) was created using a normal distribution with a mean of 120 and a variance of 50 for all regions. Lastly, cost C for the different states (corresponding to the cost of water) was obtained using a normal distribution with a mean of 400 and a variance of 100 for all regions.

TABLE 2
SIMULATION DETAILS FOR REGIONS DATA

Region	A		B		C	
	Mean	Variance	Mean	Variance	Mean	Variance
A	1300	500	120	50	400	100
B	1300	500	120	50	400	100
C	1300	500	120	50	400	100
D	600	200	120	50	400	100
E	600	200	120	50	400	100

Let us now focus on the cost of A. The intention is to assess if the cost of A is the same over all the regions. We recommend running an ANOVA comparing cost A between the five different regions. Students can see in this specific example that the cost of A over the regions is 'not the same' (i.e., that they are not statistically significant equal with a significance level of 0.05). Therefore, the example guides them towards further

establishing which of these regions are statistically different under the same statistical significance. Therefore, students are required to perform multiple pairwise tests to get an insight on which pairs of regions are statistically indiscernible (i.e., which we fail to reject the hypothesis that the costs are equal). Thus, this example shines a light on the fact that region 4 and 5 differ significantly from the rest of the regions, but we fail to reject that they are different between themselves. Hence, these findings will be in-line with the simulated data, which could be a consequence of the fact that regions 4 and 5 have subsidized seed costs. At this point, it is important to emphasize that the latter statement cannot be assessed using statistics. The causality of such subsidized system would have to be put in context (e.g., other regions do not get any financial aid, and that there is proper accountability of all variables that can affect the cost A).

C. Part III

The example is further extended to look at all the costs and their contribution to the total cost – see Table 3. Here, states are considered individually, but factors' values A, B, and C are kept the same (as displayed in Table 2). For the models created in this section, the corresponding data for each model is created following this naming convention VAR_Data. Now, we consider two models created as follows. Both Model 1 and Model 2 have an error term with a Gaussian distribution with a mean of 0 and a variance of 150 (M1_Data and M2_Data, respectively). Further, we attributed weight to each factor to be added to the total cost – see Table 3.

TABLE 3
SIMULATION DETAILS FOR LINEAR MODEL 1 AND MODEL 2 DATA

Data	Error		Weights		
	Mean	Variance	A	B	C
M1	0	150	0.70	0.15	0.15
M2	0	150	0.95	0.025	0.025

For Model 1, the weight given to A is of 0.70, whereas B and C have a weight of 0.15 each. The error was added to the total cost computed as the weighted sum of the costs. All the results can be seen in reference [15]. M1 is considered to be a linear model describing the total cost dependency on the sum of each of the costs A, B, and C.

In the second model (M2_Data), the weight attributed to A is 0.95, whereas B and C have a weight of 0.025. M2 is also a linear model describing how the total cost dependents on the sum each of the costs A, B, and C. These two models demonstrate the change in the coefficient that affects the variable costs considered. The goal of considering these two models is to show the powerfulness in modeling considerable small contributions by some variables despite a considerable amount of error being added to the final cost. That said, we have constructed the examples to conform with assumptions of linearity and normality of the noise. This is important to be emphasized to the students to allow them to connect with the theory, and then move towards the case where some of the assumptions do not necessarily hold.

Subsequently, let us consider two new models that violate the normality and linearity assumptions. Let us consider Model 3 that is violating the Gaussian distribution of error with an error uniformly distributed, randomly given to each state from -25 to 250 (M3_Data). The weights of the different costs are as follows: 0.7 for cost A, whereas cost B and C have weights equal to 0.15 - see Table 4.

TABLE 4
SIMULATION DETAILS FOR LINEAR MODEL 3 DATA

Data	Error	Weights		
	Range	A	B	C
M3	[-25, 250]	0.70	0.15	0.15

Next, a linear model was generated, which we refer to as M3, with the total cost depends on the sum of cost A, B, and C. The Residual vs. Fitted chart seen in M3 can be compared to the Residual vs. Fitted chart of M1, to shine a light on the violation simulated - see reference [15].

Next, we consider a model where there is a violation of the linearity of the model by applying different weights to the states. Specifically, in Model 4, the first 10 states have weights for A, B, and C equal to 0.95, 0.025, and 0.025, respectively. The next 10 states have weights for the costs A, B, and C of 0.70, 0.15, 0.15, respectively. Then, the next 10 states have weights of 0.34, 0.33, and 0.33 for costs A, B and C. Finally, the last 20 states have weights for costs A, B, and C equal to 0.60, 0.20, and 0.20, respectively. The error has a Gaussian distribution in Model 4 with a mean of 0 and a variance of 150 – see Table 5. We considered a linear model, referred to M4, generated with the total cost being dependent on the sum of cost A, B, and C.

TABLE 5
SIMULATION DETAILS FOR LINEAR MODEL 4 DATA

States	Error		Weights		
	Mean	Variance	A	B	C
1-10	0	150	0.95	0.025	0.025
11-20	0	150	0.70	0.15	0.15
21-30	0	150	0.34	0.33	0.33
31-50	0	150	0.60	0.20	0.20

D. Example's description in-line with students' profiles

The previous example can be customized to fit better the student audience. This overarching example allows for navigation through different topics taught in the course. Further, it shines light in the intention of the course to be connecting the dots through the topics taught; thus, allowing students to see the bigger picture. Limitations and assumptions are also highlighted through the example of the experiments conducted.

The walkthrough of how the data was generated is provided, and the key aspects were left in terms of variables to allow for the example to be molded into different scenarios. Although the example is, in general, based on a plant, the example provided allows for small modifications that enable it to be applied or adapted into other scenarios. We will walk through different ways of modifying the example in relation to catering to the profiles designed previously in the manuscript. Recall that the three main students' profiles mentioned are supply chain and logistics, data science, and academic profiles.

In the supply chain and logistics profile, there are many scenarios that can be used by adapting the example provided. For example, consider the production lines of a product (e.g., face masks). Using the face masks example, in Part I, we could consider the selling price of the mask in two different stages (e.g., before COVID-19 and during the pandemic in 2020) and

comparing it with some expected prices. In Part II, data can be simulated in relation to the cost of the face masks in different regions. Part III can consider data simulated to capture the cost of the raw materials, manufacturing costs, and shipping, which would be denoted by costs A, B, and C, respectively. Further, the different modalities of the data linearity and normality (as well as their lack of) can be explored within this example.

Alternatively, for the profile of data science, scenarios can be applied to the hands-on data example to be able to cater to this student profile. For instance, the stock market. In Part I, one stock price can be expected with a high, medium, and low cost, and then compared with data from two different periods (e.g., first 50 days of each). Part II would consist in generating the stock price in terms of group data, grouped based on season possible, and then comparing the costs throughout the seasons. Lastly, in Part III, the costs A, B, C can be considered as three separate stocks whose price is recorded in 50 consecutive business days and then summed to a total cost. The linear models can then be applied and following the reasoning provided in the previous example.

In the case of the academic profile, the hands-on example can be molded into an example that relates closely to this profile. In our specific case, we provided an example in relation to industrial hemp, with costs of seed, fertilizer cost, and water cost. This specific application is closely related in academia as the data was extracted from a paper that analyzed the industrial hemp potential and focused on the agricultural costs related to it. While simulated data was used to further develop the example, it was based on a research paper which can have further applications in the future when more data is collected about the plant. On another note, a hands-on example related to research within the pharmaceutical or medical field concerning COVID-19 can be used to cater to this profile. For example, in Part I, the price of a COVID-19 vaccine can be expected with a high, medium, and low cost, which can then be compared with two data sets collected one year apart from each other. Further, in Part II, the price of the COVID-19 vaccine can be compared between grouped data based on region from 50 countries. Lastly, in Part III, the cost producing the vaccine in each of the 50 countries can be broken down into three components: cost A, B, and C, which can be the cost of raw materials, manufacturing, and distribution, respectively. The linear models in Part III can be applied and followed based on the example provided.

VI. Supplementary Material

To help the reader reusing the examples provided in the main manuscript, and accessing a sample schedule for the revised curriculum, we made these materials available in [15].

VII. Conclusion

In conclusion, we conducted a study and a reflection on the curriculum design of a course often termed "Statistical Analysis" - an undergraduate introductory course to data science. The course objective was to equip the students with statistical methodologies that allow them to understand, process, extract, visualize, and communicate statistical evidence obtained. The designed course using system engineering tools is a stepping stone for other data science courses, as well as the social sciences or engineering courses that deal with data and seek to obtain a domain-specific technological or societal solution. This course intended to enhance students' awareness of limitations and some major methodological flaws. By designing a course that is in line with the data era, we can provide students with invaluable skills for tomorrow's job market.

Three main students' profiles were considered as a guideline for the course design (i.e., supply chain and logistics, data science, and academic profiles). Case studies and examples that cater to these specific career pathways are created with the intention of motivating and encouraging students throughout the course. Therefore, by crafting these problems to different cohorts of students, the students are more likely to pose relevant questions that can be statistically tested, for which they have to gather and organize the required data, and ultimately, translate the results obtained to a broader audience. Furthermore, the course is proposed to be set up as a hybrid between (traditional or virtual) classroom lectures and (hands-on) laboratories. By switching between these two outlets when it is most applicable, the course can provide an engaging hands-on experience for the students.

In particular, we designed an overarching hands-on example that was able to illustrate how many of the topics covered in the curriculum can be integrated into a transversal example and further be customized to the different students' profiles. Further, the example sheds light on the intention of the course to navigate through the topics and connect the dots to be able to see the bigger picture. Aside from the main concepts being demonstrated in the example, limitations and assumptions are also integrated into the problem. We provide different ways of modifying the example in relation to catering to the different students' profiles.

Although the designed course integrates practices being conducted in other universities around the United States, the course design is mainly designed catering for the data science courses in Rensselaer Polytechnic Institute. Alterations must be done to the suggested course design to best fit the courses in other universities. Furthermore, the course topics and material should be constantly altered to reflect the new and upcoming skills and trends within the market, ensuring that the course is in line with the evolving field of data science and properly prepares the students for tomorrow's job markets. The examples and material should reflect such advancements and be revisited every year.

References

[1] E. DePoy and L. N. Gitlin, "Statistical Analysis for Experimental-Type Designs," in Introduction to Research, Elsevier, 2016, pp. 282–310.
[2] "Criteria for Accrediting Engineering Programs, 2019 – 2020 | ABET." [Online]. Available: https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2019-2020/.
[3] "Industrial Engineers: Jobs, Career, Salary and Education Information." [Online]. Available: https://collegegrad.com/careers/industrial-engineers.
[4] A. Thomas, "Job Trends For Data Scientists In The Next 5 Years," Analytics India Magazine, 2020.
[5] A. Adhikari, I. Biswas, and A. Bisi, "Case Article-ABCtronics: Manufacturing, Quality Control, and Client Interfaces," INFORMS Trans. Educ., vol. 17, no. 1, pp. 20–25, 2016.
[6] R. P. Suresh, "Teaching Statistics in Management Courses in India," 2002.
[7] W. C. Parr and M. A. Smith, "Developing case-based business statistics courses," Am. Stat., vol. 52, no. 4, pp. 330–337, 1998.

[8]P. Suanpan, P. Petocz, and K. Walter, “Student Attitudes to Learning Business Statistics: Comparison of Online and Traditional Methods,” Educational Technology & Society, 2004.
[9]K. Liu, D. Klabjan, D. Shmoys, and J. Sokol, “Analytics Education - INFORMS,” INFORMS Analytics Education, 2015. [Online]. Available: https://www.informs.org/ORMS-Today/Public-Articles/October-Volume-43-Number-5/Analytics-Education. [Accessed: 01-Jun-2020].
[10] T. R. Makin and J. J. O. De Xivry, “Ten common statistical mistakes to watch out for when writing or reviewing a manuscript,” Elife,, Oct. 2019.
[11] T. C. Hesterberg, “What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum,” Am. Stat., vol. 69, no. 4, pp. 371–386, Oct. 2015.
[12] B. Efron and R. J. Tibshirani, “An Introduction to the Bootstrap,” 2000.
[13] C. O’Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, 2016.
[14] A. G. Duque Schumacher, S. Pequito, and J. Pazour, “Industrial hemp fiber: A sustainable and economical alternative to cotton,” J. Clean. Prod., vol. 268, p. 122180, Sep. 2020.
[15] Online support material at GitHub [Online]. Available: https://github.com/anagabrieladuques/IEEE-supplementary-material
[16] Biggs, J. (1996). Enhancing teaching through constructive alignment. Higher education, 32(3), 347-364.
[17] Cohen, S.A. (1987). 'Instructional alignment: Searching for a magic bullet', Educational Researcher 16(8), 16--20

Ana G. Duque Schumacher was born in Cali, Colombia in 1997. She received a B.S. in Industrial and Management Engineering from Rensselaer Polytechnic Institute in Troy, NY, USA in May 2020.

She currently is a Research Assistant in Rensselaer Polytechnic Institute in Troy, NY, USA. In 2019 she was an Information Technology Engineering Analyst Intern at The TJX Companies. She conducted undergraduate research in the Department of Industrial and Systems Engineering at Rensselaer Polytechnic Institute from 2018 to 2020. As an author of the paper “Industrial hemp fiber: A sustainable and economical alternative to cotton”, her research interest includes application and integration of sustainable fibers in the textile industry. Her research interest include data science, supply chain and logistics, and sustainable development.

Ms. Duque’s awards include the Class of 1957 Spectrum Award in 2020, Ray Palmer Baker Award in 2020, P.E. Givens Diversity Scholarship in 2019 and the Founders Award of Excellence in 2018.

Sérgio Pequito (S'09-M'14-SM’19) is an assistant professor in the Delft Center for Systems & Control that is part of the Mechanical, Maritime and Materials Engineering faculty at Delft University of Technology. Pequito's research consists of understanding the global qualitative behavior of large-scale systems from their structural or parametric descriptions and provides a rigorous framework for the design, analysis, optimization, and control of large scale systems. Currently, his interests span to neuroscience and biomedicine, where dynamical systems and control theoretic tools can be leveraged to develop new analysis tools for brain dynamics towards effective personalized medicine and improve brain-computer and brain-machine-brain interfaces. Pequito was awarded the best student paper finalist in the 48th IEEE Conference on Decision and Control (2009) and the 2016 O. Hugo Schuck Award in the Theory Category by the American Automatic Control Council.

Joana Rosado Coelho is a data scientist who has recently taught several Statistical and Data Science courses at Rensselaer Polytechnic Institute, NY. Joana is a Ph.D. in Bioengineering from MIT-Portugal program and Instituto Superior Técnico, Lisbon, Portugal and M.Sc. in Biomedical Engineering from Instituto Superior Técnico.

In the past, Joana worked in the Industry, for the financial sector where she applied text mining and natural language processing tools for automatic information retrieval. Joana was also involved in a start-up company, leading both information technology and data analysis teams and supervising all the clinical and laboratory data of the company. Joana has also a broad past experience in the energy, biomedical, and healthcare sectors. Her primary interests are in Data Science, Machine Learning, Natural Language Processing, Technology Innovation and Biomedical Engineering.

Joana was awarded the best Student of the Technical University of Lisbon and the best student of Instituto Superior Técnico in multiple years.