Application of Statistics Training to Real-World Contexts: Issues Related to Working as Data Analysts Outside Academe

H. Paul LeBlanc, III
The University of Texas at San Antonio

David A. Cortez
The University of Texas at San Antonio

Leslie E. Doss
The University of Texas at San Antonio

Through I-Corps™ customer discovery interviews (NSF Award 1925391), the authors determined that early and mid-career data analysts would be positively benefitted by the development and commercialization of an interactive software tool designed to assist in the selection of statistical tests for their real-world applications. The primary advantage addressed with this innovation is the concomitant reduction in time spent by data analysts in training and/or researching which statistical method to employ for a specific application. This paper details the development of the Stat Tree™ software prototype to accomplish those goals.

Keywords: data analysts, statistics training, software prototyping

INTRODUCTION

Stat Tree™ was developed by the first author as a tool to teach undergraduate students how to select an appropriate statistical test for a given hypothesis through a series of simple questions. The purpose of the initial development was to provide a new way of teaching the process of test selection. Originally, a noninteractive statistics decision tree was created and distributed by the first author to students enrolled in an introductory undergraduate social science research methods class beginning Fall 2001, based upon the University of Michigan Institute for Social Research *Guide* (Andrews, Klem, Davidson, O'Malley, & Rodgers, 1981, see References used in the development of the *Stat Tree*™ prototype, below) and revised for an introductory level course. In 2014, the author began developing the course for online delivery, and the first *Stat Tree*™ prototype, the web-based interactive *Statistics Decision Tree* was built by the first author utilizing *Quandary version 2* (Arneil & Holmes, n.d.).

The development of the first version of the *Statistics Decision Tree* was publicly demonstrated on May 2015 (LeBlanc, 2015), and implemented in a course taught by the first author during the Fall 2015 semester. Successive revisions were made over the next several semesters by the first author, including a more advanced version specifically targeted toward students in a graduate quantitative research methods course.

The *Advanced Statistics Decision Tree* was first implemented in a graduate level course by the first author during the Spring 2017 semester. This graduate-level version was first copyrighted using the Creative Commons Attribution-NonCommercial 4.0 International License on November 26, 2017.

This *Advanced Statistics Decision Tree* version was recommended by the UTSA Office of Commercialization and Innovation for the National Science Foundation (NSF) sponsored Southwest Super-regional *I-Corps*™ training program Summer 2018, in which the current authors participated. Thirty-three potential customers for a potential commercialized *Stat Tree*™ program were interviewed 16 different organizations throughout the state of Texas (primarily in the medical research industry in Houston and San Antonio, 48.5%). The outcome of the training was the development of a proposal for participation in the national NSF *I-Corps*™ program. As part of the NSF grant proposal process, the UTSA Office of Commercialization and Innovation submitted an application for U.S. Copyright protection (case number: 1-8253973941).

The goal of this current paper is to report on the outcomes of the customer discovery process related to the *Stat Tree*™ *I-Corps*™ project. To accomplish this goal, the authors will describe the market research associated with the potential commercialization of *Stat Tree*™*,* the intended audience and societal need for the software, and the proposed innovation and revision to the software following demonstrated potential customer need.

BACKGROUND

According to data compiled by the U.S. Bureau of Labor Statistics, job growth in fields utilizing data analysis between 2018 and 2028 will average 13.1%. Currently (May 2018 figures, updated March 29, 2019), these occupations employ 3.62 million workers. By 2028, the estimated workforce in these occupations in the U.S. will grow to 4.09 million workers, or a change of almost half a million new workers. According to Miller and Hughes (2017), significant job growth in specific data analytic fields grew by double-digit figures in 2016 alone: a) clinical data analysis, +54%; b) data science, +40%, and; c) quantitative data analysis, 38%. The field of data analysis will continue to grow at double digit numbers and outside its current industries into new sectors. Future markets may need analysts with little formal training in statistics.

The potential data analysts' market for software used to assist in statistical test selection, when segmented down to medical research only, demonstrated significant current availability and growth. The *Stat Tree*™ team investigated (as part of the *I-Corps*™ program), the Total Available Market (TAM), the Serviceable Available Market (SAM), and the Serviceable Obtainable Market (SOM) based upon data analysis labor force projections (described above). Focusing just on medical research, the TAM, comprising all medical research facilities in the United States, including pharmaceutical and medical device research, generated $171.8B in 2016, with a rise in research and development during 2013-2016 of 20.6% annually. According to Liu (2019), in 2017, the distribution software market was valued at $7.37B annually. The medical software market, worldwide, is predicted to grow at an annual rate of 5.7% over the next 5 years. The reachable market (SAM, more closely aligned with the I-Corps customer discovery), would include hospital and clinic-based research facilities, as well as contract research organizations (CROs), academic institutions, and non-profit organizations. The target market (SOM) for *Stat Tree*™ would fit within that reachable market. *Stat Tree*™'s market will expand to a wide variety of commercial applications to include financial markets, insurance, energy, etc.

As *Stat Tree*™ was originally designed to meet training needs in the education sector. Austin, Biss, and Wright, (2010) found that online study tools are preferred by students. Allen, Fielding, Westermann, and Lafratta (2021) found that students need assistance to reliably select appropriate statistical tests. Further investigation was required to determine if such a tool would be of use to professional data analysts working outside academe. Given market trends, noted above, determining the market viability of *Stat Tree*™ would depend upon the outcomes of customer discovery interviews. The specific requirements of the NSF *I-Corps*™ program included determining the needs of the target market without discussing or demonstrating the prototype. To accomplish those goals, the following research questions were posed:

RQ$_1$ *What are the job functions of data analysts working outside of academia?*

RQ$_2$ *What do data analysts working outside of academia find to be their most pressing needs related to selecting statistical tests for analysis for their jobs?*

RQ$_3$ *Would potential customers purchase and utilize a software tool for selecting statistical tests?*

METHODS

This study (IRB #19-157) was approved by the local IRB on January 17, 2019. To answer the first two research questions, formal interviews were conducted with participants face-to-face at their location (n = 112), via Skype (n = 8), or via telephone (n = 8). Travel to locations was supported through funding by the National Science Foundation (LeBlanc, Cortez, & Doss, 2019, see Figure 1). In total, 128 individuals from 35 different organizations were interviewed (see Table 1 below). The interview protocols were developed and revised to meet goals of the *I-Corps*™ program on a per week basis, based on the results of the previous week's training and interview outcomes. The interview protocols are presented in the Appendix.

FIGURE 1
MAP OF PHYSICAL LOCATIONS TRAVELLED TO FOR INTERVIEWS, EXCLUDING HAWAI'I

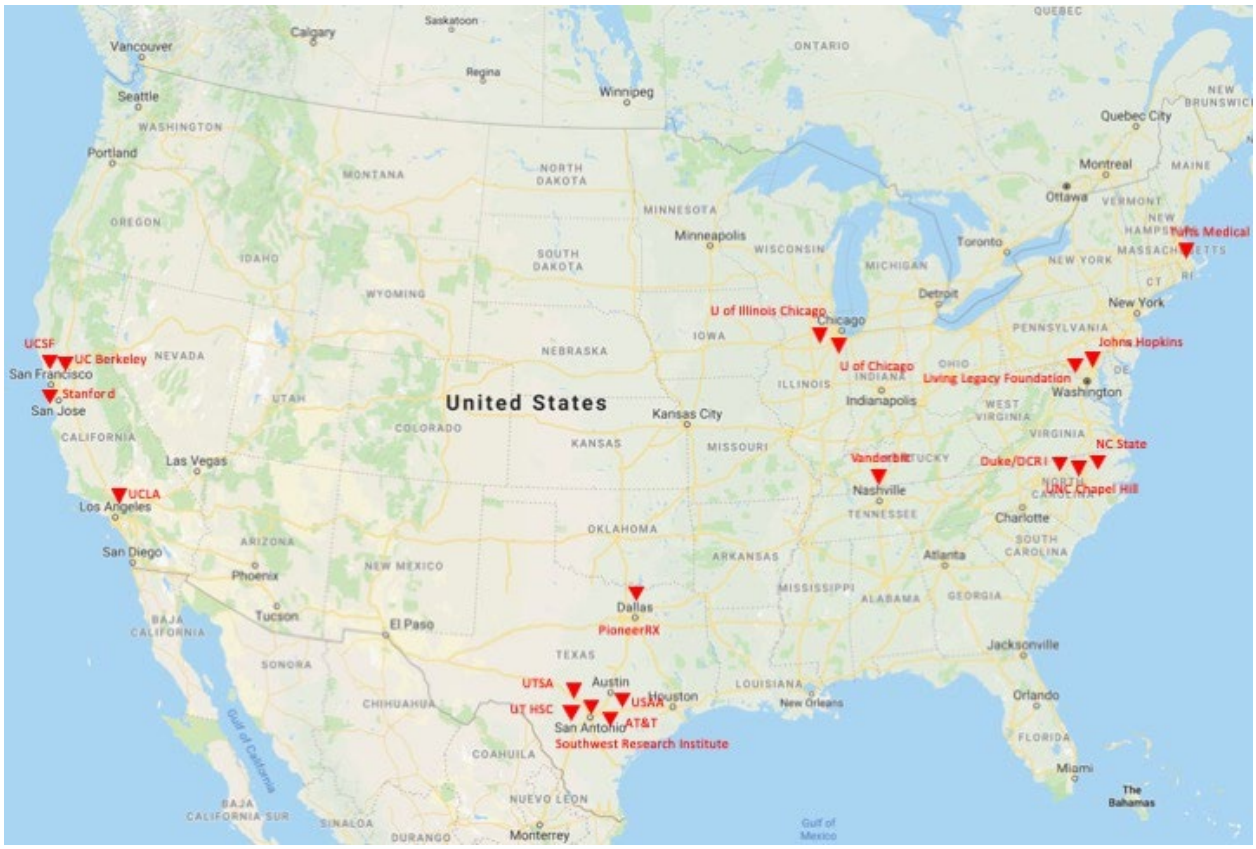


TABLE 1
ORGANIZATIONS AND NUMBER OF INTERVIEWS CONDUCTED DURING NSF *I-CORPS*™ PARTICIPATION

Organization	No.
Amazon	1
AT&T	1
Ascension Health	1
Duke Clinical Research Institute	9
Duke University Global Health Inst, School of Medicine, School of Nursing, SSRI	16
John Hopkins Bloomberg SPH, School of Medicine	10
North Carolina State University, Dept of Biomathematics, Dept of Statistics	7
National Institutes of Health	1
Pioneer RX	2
RTI International	1
Southwest Research Institute	1
St. Jude Children's Research Hospital	1
St. Thomas West Hospital	1
Stanford University, School of Medicine	6
Steve Dossin Consultants, LLC	1
The Living Legacy Foundation	1
Tufts University Medical Center	10
U.S. Government/Unspecified	1
UC Berkeley, School of Public Health	2
UCLA, Fielding School of Public Health	15
UCSF, School of Medicine	1
UNC Chapel Hill, School of Medicine	3
University of Chicago, School of Medicine	6
University of Illinois Chicago, School of Public Health	4
University of Hawaii, School of Medicine	3
UTSA	12
UT Health San Antonio	1
USAA	1
Vanderbilt University Medical Center	9
Total	128

The order of travel was as follows: 1) Nashville, TN; 2) Raleigh-Durham, NC; 3) Baltimore, MD; 4) Boston, MA; 5) Dallas, TX; 6) Los Angeles, CA; 7) San Francisco Bay Area, CA; 8) Raleigh-Durham, NC (second round); 9) Chicago, IL; and 10) Honolulu, HI. The location of interviews was selected among top medical researcher centers in the U.S. (Waber.org, 2019; Top Medical Coding Schools, 2019). Interviews were conducted by all three authors with the first author as lead interviewer for most face-to-face interviews with the second author taking notes and follow-up questions and the third author conducting the majority of the telephone and Skype interviews.

To answer the third research question, students enrolled in classes that required statistical analysis were presented with a video demonstration of the *Stat Tree*™ prototype and a survey of 5 questions regarding usefulness of the software and willingness to purchase the software upon availability.

Respondents

The respondents of the pilot study as part of the Southwest Super-regional *I-Corps*™ training program (LeBlanc, Cortez, & Doss, 2018) included: a) Data analysts, developers, engineers, scientists (n = 8); b) Clinical Research Project directors, managers or coordinators (n = 7); c) Doctoral students (n = 6); d)

Statisticians (n = 3); e) Cybersecurity analysts (n = 2); f) Faculty (n = 2); g) Other scientists (n = 3); and h) Senior company officers (n = 2). The pilot study informed the target subject pool for the present study as well as the development of the interview protocol.

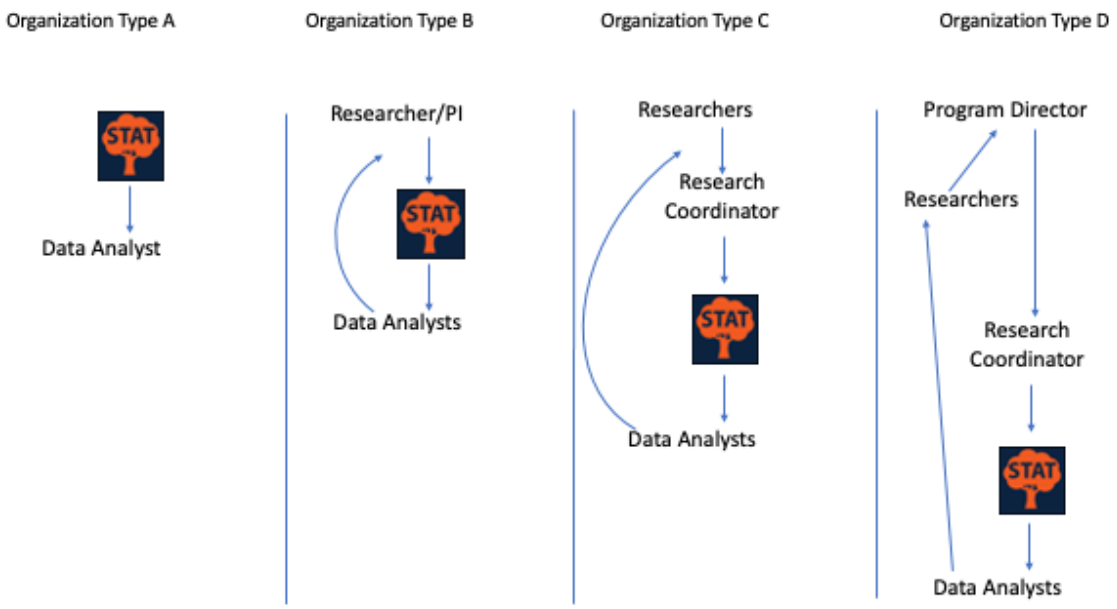
For the present study, the National Science Foundation requested that the authors more narrowly focus the interviews to data analysts/scientists and their supervisors working within the medical research field. The participants for the current study included: a) Statisticians, biostatisticians, and data analysts (n = 47); b) Research scientists and faculty (n = 28); c) Clinical research project directors, managers or coordinators (n = 14); d) Research assistants (n = 12); e) Students (n = 12); f) Directors of research (n = 10); and g) Systems analysts (n = 5). All participants in the study utilized statistics and statistical analysis as part of their job or study. Students in the survey study were undergraduate students at a large, southwestern research extensive, doctoral granting university (N = 35).

RESULTS

The major findings of the customer discovery interviews (RQ_1 and RQ_2) included:

1. The structure of quantitative medical research teams and job functions within those teams varies greatly between organizations represented in the customer discovery interviews (see Figure 2);
2. Job functions may be segregated or combined between or within individuals at the organization;
3. Job function segregation influences the degree to which research team members experience obstacles in quantitative data analysis, and;
4. The biggest obstacles experienced by data analysts, include: a) “dirty” data provided by researchers requiring their services; b) poor research design skills of researchers requiring their services; c) need for “on-the-job” training time to contextualize research problems applied within the industry in which the data analysts are employed; d) need to reduce time for selecting and analyzing statistics to accomplish more simultaneous research projects assigned, and; e) need to increase confidence in utilization of statistical techniques. These particular obstacles were more acute with early and mid-career data analysts compared to experienced data analysts, and among data analysts in organizations where job functions were more highly segregated. Data collected during the customer discovery phase allowed the team to calculate: a) an 87.5% reduction in the amount of time needed to find an appropriate statistical test for a given research question using an interactive statistics decision tree tool, and b) a 23.5% reduction in overall research production time using an interactive statistics decision tree tool.

FIGURE 2
JOB FUNCTION FOR DATA ANALYSTS BY ORGANIZATIONAL STRUCTURE



Some early key insights from the interviews were that researchers expected data analysts to be thorough and quick. However, newly hired data analysts were in need of training in applied statistical analysis. More specifically, the biggest need was that data analysts need more training on developing a method, and communicating proper statistical analysis, to the analysis and interpretation of results and even presenting the statistics. It takes time for data analysts to be proficient in their jobs in data analysis and need further professional development. Data analysts turned to Google, GitHub, or Stack Exchange to find answers to questions such as "What statistical test do I run?" The significant problem for data analysts involves receiving bad data from clinicians and researchers and the poor understanding they experience with clinicians and researchers of issues related to data analysis. Part of their job as data analysts involves making sure they get the statistics right. However, without proper communication in the design phase between the clinicians and research scientists and the data analysts, the data analysts job became more challenging.

Later key insights were that data analysts spend time educating clients (physicians/clinicians) about the statistical process, assuring that the right questions are being asked and providing intelligent responses more than doing the actual statistics. For data analysts, constant communication between them and their clients is of paramount importance. Early career data analysts want to be correct when selecting the statistical test to spare themselves from the reputational embarrassment. However, making sure the right statistical method is being used, then gaining the confidence that the selected statistical method is still the best solution to use can take two to three weeks. Often, clinicians and researchers collect data with very little knowledge about how to conduct statistical analysis resulting in data that is inappropriately formulated or measured for the tests needed for that analysis. For data analysts, setbacks occur when Principal Investigators lack knowledge about statistics. Both researchers and data analysts stated a desire for a "push-button solution."

Interviews also revealed that professional data analysts were more likely to be utilizing R, SAS, and Stata (in that order), with the majority (over 60%) utilizing the R Program. R is an open-source programming language specifically designed for statistical analysis. As an open-source solution, packages (small software programs) designed to run particular statistical tests may be written by data scientists and are typically made available to other data analysts and the general public through the Comprehensive R Archive Network (CRAN). The most common response to the interview question, "where do you go to seek answers for your data analysis questions?" was a general Google search on the Internet. The locations

for answers to statistical programming questions varied, including StackExchange.com, GitHub.com, R-Bloggers.com, and university or organizational websites. When questioned about vetting information from Internet searches for answers to statistical questions, respondents indicated that verifying information was a time-consuming task.

Survey results indicated strong perception of usefulness of *Stat Tree*™ for selecting statistical tests for data analysis, with 60.0% of respondents reporting the software as extremely useful (n = 21), 74.3% of respondents reporting the software as at least moderately useful (n = 26, cumulative), and 91.4 % of respondents reporting the software as at least slightly useful (n = 32, cumulative). Only three students (8.6%) reported the software as neither useful nor useless (n = 2), or extremely useless (n = 1).

Students were asked to indicate willingness to purchase *Stat Tree*™, purchase price and license type. The results of these questions are presented in Table 2, below.

TABLE 2
PURCHASE DECISIONS BY STUDENTS FOR *STAT TREE*™ SOFTWARE

How much would you be willing to spend to lease it for 4 months (1 semester)?					
		N	%	Valid %	Cumulative %
Valid	$30	17	48.6	48.6	48.6
	$50	10	28.6	28.6	77.1
	$80	2	5.7	5.7	**82.9**
	Does not apply	6	17.1	17.1	100.0
	Total	35	100.0	100.0	
How much would you be willing to spend to lease it for 1 year (3 semesters)?					
		N	%	Valid %	Cumulative %
Valid	$50	12	**34.3**	34.3	34.3
	$100	10	**28.6**	28.6	62.9
	$150	9	**25.7**	25.7	**88.6**
	Does not apply	4	11.4	11.4	100.0
	Total	35	100.0	100.0	
How much would you be willing to spend to purchase a perpetual license?					
		N	%	Valid %	Cumulative %
Valid	$200	10	**28.6**	28.6	28.6
	$250	6	**17.1**	17.1	45.7
	$300	6	**17.1**	17.1	**62.9**
	Does not apply	13	37.1	37.1	100.0
	Total	35	100.0	100.0	

Students were also asked to select all statistical packages (out of SPSS, SAS/JMP, Stata, and R) they would like to see covered by the software. In descending order, the students requested coverage of SPSS (60.0%), SAS/JMP (48.6%), Stata (45.7%), and R (37.1%).

DISCUSSION

The specific goal of the *Stat Tree*™ project was the development and implementation of a software tool, *Stat Tree*™, which enables analysts in multiple fields to find the most appropriate statistical test for their given application in the shortest amount of time possible, demonstrate the use of the test, the outcome and interpretation of that test, and provide the appropriate reporting of test results. The current prototype for *Stat Tree*™ meets those goals by covering 28 different bivariate and multivariate, parametric and non-parametric tests. However, the current prototype only demonstrates statistical analysis using SPSS

statistical software. This limitation occurred because the prototype was developed originally for the use of students in classes taught by the first author. However, the current study revealed that the majority of data analysts in the field use other software for their statistical analysis in order of likelihood: *R*, SAS, Stata, SPSS, and other (Python, Julia, etc), with over 60% of data analysts interviewed having used *R*. The only data analysts using Python were those dealing with "big data" (see Latta, 2020), typically greater than a terabyte new data per day, including systems analysts. Julia, a relatively new programming language designed for parallel processing of "big data," was used by two researchers among those we interviewed.

Recent review of market trends with these statistical software programs reveals that the number of data jobs is higher for analysts that can program in Python and R compared to SAS, SPSS and Stata, in that order (Muenchen, 2019). However, the percent in change in available jobs was flat for SPSS and SAS, while it was rising for Stata, R, and Python, in that order (Muenchen, 2019). In terms of scholarly article publication, SPSS is more frequently cited in the methods section of quantitative studies compared to R, SAS, and Stata, in that order (Muenchen, 2019). However, these programs are trending downward in citations in favor of software that specializes in AI/ML (Artificial Intelligence/Machine Language, Muenchen, 2019). The most significant change is with the downward trend of usage with SPSS which has been the market leader for many years (Muenchen, 2019). These trends suggest that future development of *Stat Tree* should take other statistical packages into consideration.

Although the goal of the current study was to uncover the needs of current data analysts and research project leaders (data analysts' clients) as they pertain to job functions and potential tools for job completion, the study was limited further by the goals of the NSF *I-Corps*™ program which was also focused on training innovators in the academic space to become entrepreneurs in the commercialization of their ideas. As such, much of the interview time was spent trying to gauge the potential market for a prototype software package. However, the *I-Corps*™ program restricted participants from demonstrating the prototype to potential customers. The authors did engage in a small sample pilot survey in parallel with data collection for this project. The small sample pilot survey presented participants with a short video demonstration of the prototype and asked a series of questions associated with perceived usefulness of the software and likelihood of purchase. A future study could expand the sample beyond students to data analysts working in the field.

ACKNOWLEDGEMENT

Funding for data collection of this project was provided through a grant from the National Science Foundation.

REFERENCES

Allen, P.J., Fielding, J.L., Westermann, A.H., & Lafratta, A.M. (2021). Training structural awareness with StatHand: A 1 year follow-up [Online]. *Teaching of Psychology*. https://doi.org/10.1177/0098628320985080

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Andrews, F.M., Klem, L., Davidson, T.N., O'Malley, P.M., & Rodgers, W.L. (1981). *A guide for selecting statistical techniques for analyzing social science data* (2nd ed.). Ann Arbor, MI: The University of Michigan: Institute for Social Research.

Andrews, F.M., Klem, L., Davidson, T.N., O'Malley, P.M., & Rodgers, W.L. (1981). *A guide for selecting statistical techniques for analyzing social science data* (2nd ed.). Ann Arbor, MI: The University of Michigan: Institute for Social Research.

Arneil, S., & Holmes, M. (n.d.). *Quandary version 2* [computer software]. Retrieved from http://www.halfbakedsoftware.com/quandary.php

Austin, T.L., Biss, J., & Wright, C. (2010). Student use of online study tools in business communication courses. *Journal of Strategic Innovation and Sustainability*, *6*(4), 46-54.

Babbie, E. (2002). *The basics of social research* (2nd ed.). Belmont, CA: Wadsworth.
Blalock, H.M., Jr. (1979). *Social statistics* (Rev. 2nd ed.). New York: McGraw-Hill.
Child, D. (1970). *The essentials of factor analysis.* London: Holt, Rinehart, and Winston.
Christensen, L.B., & Stoup, C.M. (1991). *Introduction to statistics for the social and behavioral sciences* (2nd ed.). Pacific Grove, CA: Brooks/Cole.
Cooley, W.W., & Lohnes, P.R. (1971). *Multivariate data analysis.* New York: John Wiley & Sons.
Coolidge, F.L. (2006). *Statistics: A gentle introduction.* Thousand Oaks, CA: Sage.
Davio, K. (2017, November 15). Report: US medical health research spending on the rise, but for how long? [Online]. *American Journal of Managed Care*. Retrieved December 6, 2019, from https://www.ajmc.com/focus-of-the-week/report-us-medical-health-research-spending-on-the-rise-but-for-how-long
Edwards, A.L. (1984). *An introduction to linear regression and correlation* (2nd ed.). New York: W. H. Freeman.
Edwards, A.L. (1985). *Multiple regression and the analysis of variance and covariance* (2nd ed.). New York: W. H. Freeman.
Glass, G.V., & Stanley, J.C. (1970). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice-Hall.
Kirkpatrick, L.A., & Feeney, B.C. (2003). *A simple guide to SPSS for Windows: For versions 8.0, 9.0, 10.0, & 11.0* (rev. ed.). Belmont, CA: Thomson/Wadsworth.
Latta, M. (2020). Sensemaking and big data science used in marketing to solve social and behavioral problems. *Journal of Strategic Innovation and Sustainability*, *15*(7), 38-42.
LeBlanc, H.P., III. (2015, May). *Utilizing online interactive decision trees for teaching complex process skills.* Demonstration at the 11th annual Innovations in Online Learning Conference, San Antonio, TX.
LeBlanc, H.P., III., (PI)., Cortez, D.A., & Doss, L.E. (2018, June). *Southwest Regional NSF I-Corps: Stat Tree.* National Science Foundation, Southwest Regional Node ($3,000, Internal University support grant).
LeBlanc, H.P., III., (PI)., Cortez, D.A., & Doss, L.E. (2019, May). *I-Corps: Interactive Statistical Decision Trees for Application in Real-world Contexts.* National Science Foundation, ($50,000). NSF Award Number: 1925391.
Lehman, A., O'Rourke, N., Hatcher, L., & Stepanski, E. (2005). *JMP for basic univariate and multivariate statistics: A step-by-step guide.* Cary, NC: SAS.
Liu, S. (2019, February 18). *Size of the distribution software/application market worldwide from 2015 to 2020 (in million U.S. dollars)* [Online]. Statista.com. Retrieved December 9, 2019, from https://www.statista.com/statistics/643858/worldwide-distribution-software-market-size
Mertler, C.A., & Vannatta, R.A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Los Angeles: Pyrczak.
Miller, S., & Hughes, D. (2017). *The quant crunch: How the demand for data science skills is disrupting the job market* [Online]. Burning Glass Technologies. Retrieved November 24, 2019, from https://www.burning-glass.com/wp-content/uploads/The_Quant_Crunch.pdf
Muenchen, R.A. (2019). *The popularity of data science software* [Online]. Retrieved March 24, 2021, from http://r4stats.com/articles/popularity/
Norusis, M.J. (2005). *SPSS 13.0 statistical procedures companion.* Upper Saddle River, NJ: Prentice-Hall.
Ray, W. J. (1997). *Methods: Toward a science of behavior and experience* (5th ed.). Pacific Grove, CA: Brooks/Cole.
Reinard, J. C. (2006). *Communication research statistics*. Thousand Oaks, CA: Sage.
Rubin, R.B., Rubin, A.M., & Piele, L.J. (2006). *Communication research - Strategies and sources* (6th ed.). Belmont, CA: Thomson Wadsworth.
Smith, M.J. (1988). *Contemporary communication research methods.* Belmont, CA: Wadsworth.

Sommer, B., & Sommer, R. (2002). *A practical guide to behavioral research: Tools and techniques* (5th ed.). New York: Oxford University.
The Comprehensive R Archive Network. (n.d.). [Online]. Retrieved October 10, 2019, from https://cran.r-project.org/
The Express Wire. (2019, December 3). *Medical software market 2019 with top countries data: Market size, growth, industry trends, share, top key players analysis and forecast research* [Online]. Retrieved December 9, 2019, from https://www.theexpresswire.com/pressrelease/Medical-Software-Market-2019-With-Top-Countries-Data-Market-Size-Growth-Industry-Trends-Share-Top-Key-Players-Analysis-and-Forecast-Research_10471689
Top Medical Coding Schools. (2019). *25 Universities affiliated with top teaching schools* [Online]. Retrieved February 18, 2019 from: https://www.topmedicalcodingschools.com/25-universities-affiliated-with-top-teaching-hospitals/
U.S. Bureau of Labor Statistics. (2019). *Occupational projections and worker characteristics (2018-2028)*. Retrieved September 20, 2019, from https://www.bls.gov/emp/tables/occupational-projections-and-characteristics.htm
Vogt, W. P. (2007). *Quantitative research methods for professionals.* Boston: Pearson Education.
Vogt, W.P. (2005). *Dictionary of statistics and methodology: A nontechnical guide for the social sciences* (3rd ed.). Thousand Oaks, CA: Sage.
Wisconsin Association for Biomedical Research and Education. (2019). *Top medical research centers in the U.S.* [Online]. Retrieved February 18, 2019, from http://www.wabre.org/top-medical-research-centers-in-the-us

APPENDIX

Interview Protocol by Week

Week 1:

A. What is your job title/role?

Data analysts, statisticians, biostatisticians, systems analysts, research assistants, and students

b. How many years have you been working as a data analyst?
c. Tell me about your job functions?
d. How did you train for this function?
e. What problems did you encounter in your training?
f. Where do you go to seek answers for your data analysis questions?

Clinical research project directors, managers or coordinators, directors of research, and Research scientists and faculty

b. How many years have you been working as a research coordinator?
c. Tell me about your job functions?
d. What is your relationship with data analysts?
e. What problems did you encounter working with data analysts?
f. Do you have to train new/early career data analysts?

All participants

g. What do you see as your biggest need? Can you describe it? Why is that important?
h. What do you see as your second biggest need? Can you describe it? Why is that important?

i. Who is responsible for decisions in meeting that need (Job title/role)?
j. What question did I not ask that I should have?
k. Who else should I speak with?
l. Can I follow-up with you later?

Week 2 (new questions)

Data analysts, statisticians, biostatisticians, systems analysts, research assistants, and students

a. What problems do you encounter in your job?
b. How do you typically resolve these problems?
c. What do you see as a possible solution that is not currently available?
d. What do you gain from solving problems you encounter in your job?
e. What is keeping you from getting there?
f. What would you pay to get there?

Clinical research project directors, managers or coordinators, directors of research, and Research scientists and faculty

a. How do you typically resolve problems working with data analysts?
b. Where do data analysts go to meet that need?
c. What do you gain from solving problems you encounter in your job?
d. What is keeping you from getting there?
e. What would you pay to get there?

Week 3 (new questions)

Data analysts, statisticians, biostatisticians, systems analysts, research assistants, and students

a. What problems do you encounter in selecting the proper analysis for a given research question?
b. What motivates you to find the proper statistical test for a given research question?
c. How long does it take to find the proper statistical test for a given research question?
d. How does it affect you when you find the proper statistical test for a given research question?
e. Where do you look to find the proper statistical test for a given research question?
    i. If you use Google, how much time does it take to find the best answer?
    ii. If you use textbooks, how much time does it take to find the best answer?
    ii. How do you determine if the information you found is correct or credible?
f. How often does it happen that mistakes are made in choosing the correct statistical test? Or correct options for a given statistical test?
g. What does it cost if a mistake is made?
h. What would you do to reduce the likelihood of mistakes in choosing the appropriate statistical test and options?

Week 4 (new questions)

Data analysts, statisticians, biostatisticians, systems analysts, research assistants, and students

a. What is your relationship with researchers?
b. How much does that time spent finding the proper statistical test cost?
c. What would you spend to avoid that lost time?
d. What would you spend to avoid mistakes in choosing the appropriate statistical test?

Clinical research project directors, managers or coordinators, directors of research, and Research scientists and faculty

a. What would you pay to solve problems related to data analysis you encounter in your job?
b. Who makes purchasing decisions about solutions?

Final questions (no new questions weeks 5 through 7)
All participants

a. Tell me about how research work involving statistical analysis gets done.
b. Tell me a little about the projects you work on?
c. How many projects do you work on in a week?
d. How many PIs do you work within a week?
e. How long does it take you typically (ballpark) to find the correct test on a given project?
f. How long does it take for you to work on a given project?
g. Where would you prefer to purchase a solution?