Spring 2021 For Members and Friends of the NV ASA

SILVER STATE-ISTICS

Nevada Chapter of American Statistical Association

Nevada Chapter Newsletter Vol. 18 No. 2

Dear Nevada ASA Members and Friends,

Read on for information on
what our Chapter has been up to.
Please keep an eye on our website
for more details on upcoming events.

https://community.amstat.org/nevadachapter/home

Contents:

Coming Events 1
The President's Corner 1
Virtual Short Course: Introduction to Data Science, Machine Learning and Deep Learning 2
StatFest 2021 2
Careers in Data Science and Statistics: Virtual Symposium 2 2
CDSS: Virtual Symposium 3 3
NV-ASA Chapter Governance and Constitutional Convention 3
NV-ASA is Now on LinkedIn 3
NV-ASA Turns 20! 4
Joining NV-ASA! 4
NV-ASA Officers & Others 4
Software for Undergraduate and Graduate Education in Data Science and Statistics 5

COMING EVENTS

We're looking forward to a short course on Data Science, Machine Learning, and Deep Learning July 23-24 (see below); a symposium on Data Confidentiality (~September 11); and our Annual Meeting (~ October 16).

%

The President's Corner

Christopher Tong

Greetings. Let me begin by thanking all three speakers in our Careers in Data Science and Statistics Virtual Symposia series: Julia Anderson, Lissa Callahan, and Brian Karp. Each session ran half an hour past the scheduled time, illustrating the high level of interest from attendees. Recordings of the first two symposia are available at our new YouTube site: https://www.youtube.com/channel/UC-3UbP528AVylhQFaN8DzrQ (The third symposium was not recorded.)

We are also grateful to the co-sponsor of the series, the UNR Dept of Mathematics and Statisics. Prof T.Y. Tam, department chair, initiated this collaboration, and Prof Mihye Ahn managed the event registration and zoom hosting logistics. This series could not have happened without their partnership. A special thank-you is due to our former Southern Vice President, Shar Chaudhry, who helped recruit two of the speakers. We look forward to hearing your feedback on this series, which you can provide by contacting me directly, or by posting comments on our new Linked-In page (just search for Nevada Chapter of the American Statistical Association).

NV-ASA is open to brainstorming other ideas for collaboration with any Nevada educational institution with regards to data science and statistics. We have a mutual interest in equipping your students to be successful in the job market and cultivating a data literate citizenry.

The chapter is planning at least three events for the rest of the year: a short course (courtesy of the ASA Council of Chapters) on data science, machine learning, and deep learning; a symposium on data privacy and confidentiality; and of course our annual meeting. You can read more about the short course in the attached flyer. Planning for the other two events is very early at this stage, but they will likely be in the September - October time frame of the year. Σ

Elections

Positions coming open for 2022 are President, Northern Vice President, and Secretary; be thinking about whom you would like to nominate! Election results for terms beginning January 1, 2021 were announced during our Fall Annual Meeting. Members of the Executive Committee are listed at the end of this Newsletter. ∞

Virtual Short Course in July: Introduction to Data Science, Machine Learning and Deep Learning

This virtual short course will be sponsored by NV-ASA with support from the ASA Council of Chapters. It will take place in two 3.5-hour afternoon sessions on Friday and Saturday July 23 and 24. The aim is to provide an overview of using R and Python for some of the most popular machine learning and deep learning models in real-world applications in the cloud environment. Instructors will be Ming Li, a senior research scientist at Amazon, and Hui Lin, a quant researcher at Google. The background suggested is bachelor-level coursework in statistics or something similar. See the accompanying flyer for more details and a registration link. ≤

StatFest 2021

StatFest is a conference for undergraduate students organized by the ASA's Committee on Minorities in Statistics. It will take place Saturday and Sunday September 18 and 19. StatFest encourages historically underrepresented undergraduate students to consider careers and graduate studies in the statistical sciences. It features "two packed half days" of discussions from professionals, academic leaders and current graduate students. For more information, see https://community.amstat.org/cmis/events/statfest/statfest-2021. ΣΣ

Careers in Data Science and Statistics Virtual Symposium 2

Our CDSS – VS2 took place Tuesday evening March 23, featuring Lissa Callahan, Data Analyst with ADM Associates along with her ADM colleagues Alexandra Horne, Christopher McBride, and Justin Merkel, and a total of 53 participants. The theme was **Data Analytics in Energy Efficiency**. Lissa graduated from UNR in 2019 and joined ADM shortly thereafter. Her colleagues described their varied career paths as well: Alexandra coming from Geology and modeling, Chris coming from Physics and Astronomy, and Justin coming from Software Engineering.

Lissa gave us an in-depth look into a major project of interest at ADM, involving quantitative evaluation of the effectiveness of behavior modification interventions in energy efficiency. ADM's clients are utility companies. These companies are interested in evaluating the effectiveness of various "interventions" they might use to encourage their customers to be more energy efficient. For a given client a large number of customers are selected for the study. Based on a year's prior electrical billing data, the group is divided into matched control and treatment groups, with the treatment group receiving the intervention the following year. The "Matching" R software package is used for this propensity score matching (an alternative is "MatchIt" package). Energy usage is compared for the two groups for the pre- and post-intervention periods. Both summer and winter comparisons are made. The analysis is described as a "difference-in-differences" method. Linear mixed effects regression is used, using both random (energy usage) and fixed (dummy variable) effects, and controls for how weather affects energy use. Computations for this project are done using R, but other ADM analyses also utilize EXCEL, and SQL for data retrieval. Lissa described the model validation methods used; metrics often involved *t*-tests and *p*-values. One of her "take-aways" was a recommendation to develop skills at communicating such metrics and their meaning to the clients.

During the Q&A portion there were a number of questions about what skills and technologies are used at ADM and what one should be able to present in a resume or discussion during an interview. The ADM team suggested that Python is effective and versatile. Once methods are developed, translating them to R and vice-versa is not that difficult, and there are resources that can assist. Git (GitHub or GitLab) is very useful for keeping track of changes and sharing work and code with colleagues; Lissa highly recommends including detailed comments with code that you store there. One reason for that is to demonstrate your prowess to prospective associates. Another is the reality that eventually any code will need to be updated, and the comments will make it much easier to the one doing the updating (possibly even yourself) to understand what was done and why. (There is also the CRAN repository for R code.) One can learn how to use these and other software products from DataCamp.

In response to a question about what one can show prospective employers, there are of course always the courses that one has taken, but more useful are the projects that you have successfully completed. The point is these will help demonstrate that you can solve new problems; in courses most of the time is spent reviewing already-solved problems. Φ

Careers in Data Science and Statistics: Virtual Symposium 3

Our CDSS – VS3 took place Tuesday, April 27, with around 27 participants. The presenter was Brian Karp, Senior Audit Manager, Advanced Analytics & Innovation Excellence, Wells Fargo. Brian began by describing his most interesting career path, starting with accounting and travelling through numerous steps including MIS, data analyst and internal audit, data management, and now advanced analytics including AI. His comment on career opportunities with this sort of varied background was "what can't you do?".

He described Data Science as involving a deep, common body of knowledge and combination of skills. There can be a distinction between Data Analysis, which deals with historical data, and Data Science, dealing with data yet to be obtained. Basic programming skills are essential, including knowledge of common tools and data acquisition methods (see the Q&A section below). But he also emphasized the importance of skills at communication with stakeholders: learning how you can help them with their problems and relaying your solutions to them. This can be thought of as a form of "story-telling". Also important, once you have developed procedures for solving a stakeholder's problem, the procedures should be kept up-to-date and relevant; this comes under the category of change management or procedure maintenance.

One reality that Brian told us is that your first job is unlikely to be your last job. You may follow a statistical path, focusing on methodology, or a technical path, focusing on customer needs, or some mix, possibly as a business or financial analyst. He emphasized, though, the ideal of charting a career path based on your ideals – how you can use your skills for the greater good while satisfying personal and professional values, morals and objectives.

Interesting questions came up in the Q&A segment. When asked what the focus of his group was, he said that there was a major focus on risk mitigation, involving anomaly detection, data integrity, and fraud prevention. Software skills expected of a person in this field included knowledge of and some skill in SQL (for data acquisition), SAS (very commonly used), data visualization software, and Python and R. When asked whether one should be expert at all of these, he emphasized the idea that one should be able to learn what one needs to use efficiently.

For career advancement, Brian recommended acquiring a mix of technical and "soft" skills. Advanced positions can include associate (more technical) positions or consultant (more client-oriented) ones. One question involved staying relevant in techniques you're not currently using, to which he reported that he does often enroll in continuing education programs.

And regarding how to get noticed when applying for a position, his recommendation was to ensure that your resume includes of course the technical background expected but also a broad background showing an open-minded attitude. λ

NV-ASA Chapter Governance and Constitutional Convention

NV-ASA is launching an effort to review and possibly amend our Chapter Constitution, bringing certain elements in line with national ASA requirements and modifying language to incorporate current practices such as allowing email ballots and virtual meetings, formally recognizing Lifetime membership, possibly adding membership categories, and revising language throughout to include data science as well as statistics. A Chapter Governance Committee has been formed, chaired by former president Deb Stiver. It is likely that we will host a forum open to all members to discuss proposed changes. All chapter members are encouraged to submit ideas and proposals. The aim is to propose changes either prior to or during our October Annual Meeting. Φ

NV-ASA is Now on LinkedIn

We now have a LinkedIn page for our Chapter. Visit https://www.linkedin.com/company/nvasa. If you are already a LinkedIn member, you will go to our page directly. If not, you will need to join first. One feature is that our NVASA page will allow job postings. To do that, please contact our webmaster at the email address given on our website. δ

NV-ASA TURNS 20!

The Nevada Chapter of the American Statistical Association (NV-ASA) was founded in 2002. We will be celebrating our 20th anniversary next year. Brain-storming is underway for special events and celebrations, hopefully in-person as well as remote! <==>

Joining NV-ASA!

Only a minority of the people who receive this newsletter are members of the Nevada Chapter of the American Statistical Association (NV-ASA). Dues are nominal. For full-time students at NV institutions, the cost is $10 per year ($2 for student members of the national ASA). Otherwise, it is $20 per year ($10 for members of the national ASA). One can become a Life Member for $100. You can join NV-ASA when you renew your national ASA membership (or join for the first time); this can be done on-line at **www.amstat.org/chapters**.

Otherwise, whether a national ASA member or not, you can join through PayPal on our website or by contacting our Treasurer Alejandra Livingston. Any way you do it, please obtain an information form from our website, complete it, and send it to Alejandra at the address listed on the form.

Why should you join? NV-ASA events provide opportunities for networking and contact with other statisticians working in a wide variety of areas in Nevada. But in addition to that, a major reason is that your dues help support the outreach activities of the NV-ASA including the K-12 Poster Competition and Career Days. Our financial needs are not great, so long as we all pitch in our modest amounts.

NV-ASA Officers & Others

Voting officers are

President:	Christopher Tong	(2020-2021)
Past President:	Alicia Chancellor Hansen	(2020-2021)
Northern Vice President:	Glenn Waddell	(2020-2021)
Southern Vice President:	Sean Breckling	(2021-2022)
Secretary:	Charles Davis	(2020-2021)
Treasurer:	Alejandra Livingston	(2019-2022)
Chapter Representative:	Gayle Allenback	(2020-2022)

Also involved are

Webmaster:	Alicia Chancellor Hansen
Poster Competition Lead:	Elizabeth Harris
Newsletter Editor:	Charles Davis

Silver State-istics welcomes news items and letters from members and friends of the NV-ASA on matters of interest to the Chapter and the profession. Manuscript or items can be sent as a Microsoft Word document, PDF, or within an e-mail.
Silver State-istics is published by the Nevada Chapter of the American Statistical Association.
All items appearing in *Silver State-istics* are from the individuals providing them, and are not intended in any way to represent positions or opinions of any employer, government agency, client, the NV Chapter of the ASA, or the National ASA.

For contact information, go to
https://community.amstat.org/nevadachapter/home
Our address for regular mail is
NV-ASA, PO Box 3311, Sparks, NV 89432-3311

Software for Undergraduate and Graduate Education in Data Science and Statistics

Christopher Tong

Introduction

During the Q&A for Dr. Julia Anderson's presentation at our February careers symposium, a question was raised about what software should be taught in Statistics and Data Science degree programs to amplify students' marketable skillsets. The question was stratified into lower-division undergraduate, upper-division undergraduate, and graduate levels. The proposed answers from Dr. Anderson and other participants were sound and based on real-world experience. I would like to recast and extend some of that discussion here and invite responses from others. I also strongly believe that such a question needs to be addressed within a broader thought process about how computing should be integrated throughout the data science and statistics curriculum. Fortunately, the March 2021 supplemental issue of the *Journal of Statistics and Data Science Education* is a special issue devoted to this broader rethinking (https://www.tandfonline.com/toc/ujse21/29/sup1), so I need not dwell further here on those important considerations.

In this note I will group statistical analysis software into three categories:

1. Point-and-click software. Examples include JMP, Minitab, NCSS, Prism, SigmaPlot, SPSS, and Statgraphics. Note that many of these also have capabilities for script programming or developing macros, in addition to the simple graphical user interface.

2. Scripting languages/software. Examples include Python, R, and SAS, though all three have powerful capabilities that bleed into the next category. (Note: Python is a general purpose interpreted language whose statistical and machine learning capabilities have been dramatically enhanced in the last decade. Python is readily integrated into other systems and software.)

3. Statistical computing and software development. Many of us will be called on to build software to be used by others, including non-statisticians. Requirements may run the gamut from writing a simple Excel macro in VisualBasic to creating a computationally intensive R package, Python library, or SAS macro or developing standalone, executable software for enterprise use. Advanced practitioners will need to learn a compiled programming language such as C++ or Java, and interested students should be advised to seek courses from a computer science/engineering department. Advanced computing courses in algorithm analysis, numerical analysis, etc., could be valuable.

There are also non-statistical software systems that every data scientist and statistician should learn, as Dr. Anderson noted, such as SQL for databases and spreadsheet software Excel (more on that in a moment). Many positions in our field require use of "dashboard" software such as Tableau. It would be to many students' advantage to have a personal site on GitHub where they can share code that prospective employers may view.

Finally, this note focuses on statistical *analysis* software. Another discussion entirely may be had on software for experimental design, but I have no special expertise on this matter. Throughout my career I have designed studies "manually" or via simulations in a statistical scripting language. However, specialized commercial software for either classical experimental design or clinical trials design can be immensely powerful and time-saving.

How to answer the question

The best advice I can offer students and curriculum developers is to act like a statistician: gather some data relevant to your situation. The cheapest and easiest way to do this is to make a habit of reading current job advertisements.

Students should start doing this years before they graduate, and focus on the industries and geographic regions of greatest personal interest. In reading these ads, students should think about what software (and, frankly, other) skills they'd like to add to their resume, and take the initiative to learn them, even if their home department does not offer opportunities to do so. This will require taking courses from other departments, and most likely, signing up on your own for bootcamps, workshops, and short courses at conferences or online, such as through massive open online courses (MOOCs). Once you learn a software on your own like this, endeavor to keep using it on real projects, to develop deeper experience and prevent skill atrophy.

Curriculum developers should take a broader sample of job ads, in order to get in touch with, and take the pulse of, the job market for their graduates. Much of what I offer below is based on my own (somewhat selective) reading of job advertisements, and from speaking with fellow data professionals in various industries.

As Dr. Anderson noted, LinkedIn has become a tremendous resource for finding job advertisements. However, many government jobs don't get posted there. Federal government positions may be found in a single place, usajobs.gov. Of course, the national ASA maintains a job board on its website.

Elephant in the room 1: Excel

For better or worse, Microsoft Excel is by far the most commonly used software for data analytics. However, Excel encourages many bad habits that could collectively be described as promoting nonreproducible analyses and degrading data integrity; in addition, for many years many of Excel's statistical algorithms were demonstrably incorrect, and others were known to be numerically unstable. (It remains a subject of debate how extensively these have been repaired or at least improved in recent years.) Moreover, Excel is not efficient with memory usage, and not capable of opening very large data files. Ideally Excel should be completely avoided for analytics; even the most rudimentary point-and-click statistical software should be preferred. However, reality does not correspond to the ideal world. As Dr. Anderson noted, many of our collaborators and customers will send us data in Excel, and they may expect work product returned in the same format. Some may even demand that we provide Excel macros to execute any analytics procedures that we want them to use. Data scientists and statisticians must be prepared to deal with Excel, but they should also be well educated about the hazards of using it for analytics, a topic that deserves a full lecture that I won't give here.

Elephant in the room 2: SAS

This is not the place to debate the merits of different statistical software or the reasons why some are widely used in certain industries. We need to accept the facts on the ground. SAS remains an industry standard in large swaths of the pharmaceutical and finance industries, as well as in different corners of the federal government. These are some of the largest employers of statisticians in the United States. Moreover, there are strong reasons to expect SAS to continue its dominance for the foreseeable future, including its back-end architecture, interfacing with other systems like R and Python, and new features offered in the SAS Viya ecosystem. Every serious degree program that claims to prepare statisticians for the job market beyond academia must, at least as an option, provide intermediate proficiency in using SAS. Without this option, many exciting and rewarding job opportunities will remain firmly closed to their graduates. SAS also offers machine learning and other data science capabilities, and is used for this purpose by our very own Desert Research Institute's Healthy Nevada Project, for instance.

Recommendations

I agree with Dr. Anderson that Python has emerged as the most ubiquitous common language of data science. Though I am personally not a Python user, my reading of the job advertisements in data science (as well as our 2017 annual meeting keynote speaker, Dr. Ming Li) suggests that no graduate will get very far without some proficiency

in Python. **Thus, upper division undergraduate and graduate degree programs must offer students the option of learning either Python, SAS, or both.** R maintains a large and active user community, and the emergence of tidyverse has helped put the focus of statistics back on working with data, as evidenced by Hadley Wickham receiving the COPSS President's Award in 2019. If a third option could be entertained, I would make it R. (In full disclosure, I have principally been an R user for about 20 years.) Lissa Callahan's March career symposium illustrated the use of tidyverse tools along with rmarkdown with R. A recent presentation sponsored by the National Institute of Statistical Sciences (NISS) given by Prof. Mine Cetinkaya-Rundel, "Toolkit for the Modern Statistician", provides an overview: https://www.niss.org/news/what-should-be-modern-statisticians-toolkit.

Lower division undergraduate courses might consider the use of point-and-click software, especially if we want students to focus on concepts, not on coding. Many of these packages have well-established niches in different industries, providing real-world benefit. However, point-and-click software often allows users to do analyses without fully documenting everything they did, rendering the analysis potentially nonreproducible (though not nearly as badly as Excel does). Often only the "final" version of the analysis is saved as a file. There are strong scientific, statistical (e.g., inferential multiplicity), and legal (eg, 21 CFR Part 11) reasons to strongly discourage nonreproducible analyses and document all interaction with the data. For this reason, some point-and-click software options offer access to a scripting language, or they may generate an audit trail log file, capturing all the user's interactions during a session. Alternatively, an argument could be made that all STEM students should learn coding, so that even an introductory course should consider Python or R, which have the additional advantage of being freely available, unlike nearly all point-and-click alternatives. Dr. Glenn Waddell noted that for those unwilling to install R on their computers, https://repl.it/ can be used on the web, though it does not access packages.

The question regarding which approach to use for lower division courses is unsettled. The GAISE (*Guidelines for Assessment and Instruction in Statistics Education*) remains agnostic on the matter, but provides valuable discussion of how technology can be used in such courses. Ultimately the choice a specific degree program makes will reflect its educational philosophy and differentiate it from competing programs. I refrain from providing a recommendation at the lower-division undergraduate level.

Finally, Dr. Anderson adds that "students and professionals must have the forethought and drive to learn other languages as they come to the forefront. Python for example, did not have the statistical or package capabilities 5 years ago (when I had to choose a language for my dissertation), but in that short time it has surpassed R for functionality, ease of learning, and available packages that are all community supported. At the time, I chose R because it made sense, however, I now know that our Cigna servers are slowly all transitioning from R/SAS/Python to just Python so I'll be transitioning at some point. Also, the benefit with starting with statistical scripting software like Python is that it's very easily translated to other languages. I started out learning Python but was easily able to transition to R." Readers may also find the blog post by Matt Dancho and Jarrell Chalmers, with the oversimplified title "R is for Research, Python is for Production", of interest to this discussion: https://www.business-science.io/business/2021/02/18/R-is-for-research.html

This topic has generated great interest and NV-ASA is considering hosting a roundtable discussion on it. Please let us know your thoughts!

Acknowledgments

I am grateful to Julia Anderson, Sharang Chaudhry, and several NV-ASA executive committee members for helping me to improve this note.