



SILVER STATE-ISTICS

Nevada Chapter of American Statistical Association



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:

NV-ASA Turns 20!	1
The President's Corner	1
Elections	1
Spring 2021 Virtual Symposium	2
Careers in Data Science and Statistics:	
Virtual Symposium 1	2
CDSS: Virtual Symposium 2	3
K-12 Poster Competition	3
NV-ASA is Now on LinkedIn	3
Job Opening at Wells Fargo	3
Online M.S. in Business Analytics Program	
at UNR	4
Joining NV-ASA!	4
NV-ASA Officers & Others	4
When Big Data is so Messy, You Never Get to	
Analyze It	5
StatsChats: Open-door research support at	
UNR	6

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. Brainstorming is underway for special events and celebrations, hopefully in-person as well as remote! $\leq \Rightarrow$

The President's Corner

Christopher Tong

The COVID-19 pandemic continues to challenge NV-ASA in 2021 as it did in 2020. Despite this, we have an active spring planned. I want to thank Profs. Ole Forsberg and Ken Miller for participating in our spring virtual symposium last month. They provided a capsule education for those of us, like me, who are unfamiliar with an important field of applied political science that puts many core statistical concepts to work, and one in which even more statistical input might be valuable. I would also like to thank Profs. T.-Y. Tam and Mihye Ahn, of the UNR Department of Mathematics and Statistics, for collaborating with us on a forthcoming Careers in Data Science and Statistics Virtual Symposium series. You can read more about these endeavors elsewhere in these pages. One year ago, I wrote in the President's Corner that "It is particularly valuable, in my view, to bring university students (and those who teach them) into contact with practitioners in the private, public, and nonprofit sectors." I hope that the guest speakers we've brought (and will continue to bring) to chapter events continue to fulfill this promise. Finally, I welcome new chapter officer Sean Breckling as Southern Vice President, and other friends of the chapter who have been joining us for Executive Committee meetings, as well as the returning officers. Their support is most needed and appreciated.

Σ

Elections

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. Positions coming open for 2022 are President, Northern Vice President, and Secretary; be thinking about whom you would like to nominate! ∞

Spring 2021 Virtual Symposium

Our Spring 2021 Virtual Symposium on **Perspectives in Polling Methodology** took place Saturday morning, January 30. It featured presentations by two speakers, with Q&A; 37 participants attended.

We heard first from Ole Forsberg, chair of the statistics program at Knox College in Galesburg, IL, speaking on *The Polling Narrative is Wrong: ... but there is still work for us to do*. Dr. Forsberg began with several quotes from prominent news publications on why the polls were wrong: “Why were the pollsters so off?” (*The Hill*); “Once again, reality has humiliated the polling industry.” (*New York Post*); “Were The Polls Wrong? A Look At The Future Of The Polling Industry” (*National Public Radio*); “Why the polls were wrong – and will never be right again” (*The Sydney Morning Herald*); and more. With the last of these he introduced the idea of the “shy Trump voter”, an example of “social desirability bias”, a concept prominent in the ensuing discussions.

His next section was “PLTH 101”, basic concepts of polling theory. The fundamental statistical question coming out of this is “How do we ensure that the sample is representative of the population?” – but what is the population when dealing the stratified/skewed samples and shy Trump voters? Weighting the strata is a standard technique, but to do that one has to know (at least reasonably) the proportions of the “population” in each of the strata. Different polling houses use different proportions; some have reputations as being “left-leaning”, others as “right-leaning”. Dr. Forsberg gave us numerous examples illustrating their implications. And, of course, the CIs rarely, if ever, take into account the uncertainty or variability involved in “guesstimating” the right weights for the strata.


Another major issue raised by Dr. Forsberg was communicating confidence intervals. After all, a 95% CI should miss the parameter of interest 5% of the “time”, but “time” itself is a curious concept here. And then there are the challenges of communicating such matters, as we all understand. His talk is posted on our website, with references to additional readings for those interested. Dr. Forsberg also has an article “US election polls: a quick postmortem” in *Significance* v. 18 issue 1.

Next, Kenneth Miller of the Department of Political Science at UNLV presented *2020 Presidential Polling in Nevada and Nationwide: Public poll performance in the context of expectations and history*. He started by showing us a map of the US with states color-coded **Likely Biden**, **Leans Biden**, **Toss Up**, **Leans Trump**, or **Likely Trump**, based on RCP (Real Clear Politics) poll

averages shortly before the election, followed by the update following the election. He illustrated the concept of prediction confidence via 40,000 simulations, with Trump winning in the Electoral College 10% of the “time” and Biden winning 89% of the “time”. An interesting chart followed, comparing late marketing advertising allocations per market by the two campaigns with electoral college votes, sorted by Biden vs Trump average poll predictions as compiled by FiveThirtyEight.

Dr. Miller gave a number of examples from various polling houses at various times. In all of the examples Biden was forecast to receive around 50% of the vote, whereas Trump’s share was forecast to be from 41% to 49%. This leads to the next section, “The sources of survey error: Surveys are getting things largely right, but there is a sense that something is off”. Total Survey Error has three components: Statistical Error (discussed in STAT 101), Measurement Error, and Sampling Error. The first Dr. Miller describes as being least troublesome, even though often ignored in media coverage. He does point out that when aggregating results from state level polls (say), the influence of correlated sampling and measurement errors may have little effect on Confidence/Prediction Limits themselves, but can provide an illusion of greater precision. Under Measurement Error, he refers to the Shy Trump Voter syndrome; if this is actual, he expects to see differences in Trump support by survey mode. Under Sampling Error comes the challenge of identifying likely voters for sampling, always a problem, but especially so this year with the pandemic-induced changes in voting mode.

Other issues are partisan non-response perhaps related to published poll results going badly for their candidate and distrust of institutions. (One comment arising in the discussion was that the choice of the statistical model to use is potentially another source of variability.)

Finally, Dr. Miller states that “The landscape is increasingly crowded with surveys that have political, profit, or other incentives instead of an accuracy incentive”. *Caveat emptor!* This presentation is also posted on our website. 

Careers in Data Science and Statistics: Virtual Symposium 1

The first of our Virtual Symposia on Careers in Data Science and Statistics took place Thursday, February 25. The topic was **Data Analytics in Health Insurance**, presented by Dr. Julia Anderson, Senior Business Analytics Advisor for Cigna.

Dr. Anderson first described her career path starting with studies in molecular biology and leading to her Ph.D. in

Public Health with emphasis in Epidemiology and Biostatistics from UNLV and her current involvements with Cigna. She gave us an overview of her view of the health insurance industry, and then described her day-to-day projects. She works with the CHAM (Cardiometabolic High Acuity Model) group, a ~100-member team supporting ~5 client companies. The group aims to use predictive modeling to steer health industry activities from simply reacting to claims as presented toward using (client-specific) data to develop proactive interventions to improve employee health. Particular topics she discussed include attempting to predict heart conditions, analyzing situations involving high-cost claimants, and helping claimants with musculoskeletal issues.

The second part of her presentation was on “Applying for Jobs: Tips and Tricks” and related matters. She gave us advice on resume preparation, such as making sure that the AI that reads the submitted resume will “see” that it really is related to the job you’re applying for (and keep it brief – not an academic CV).

Regarding skills a successful applicant should develop, Dr. Anderson recommends fluency in Python and R (which are free and can be learned through DataCamp, which was part of her path to her present career) and EXCEL; the latter, though perhaps “old-fashioned”, is a common way of communicating datasets among various platforms. She also mentioned the Python-based Data Robot for dealing with large datasets, along with familiarity with SQL.

The scheduled portion of this symposium was from 5:00-6:00 pm; the Q&A portion lasted another half hour. There were 44 participants. It is available on YouTube: <https://www.youtube.com/watch?v=u3dapNBBpPA>

She recommends getting involved with LinkedIn. In fact, she is very willing to provide continuing support to participants, through LinkedIn. λ

CDSS – Virtual Symposium 2

The next Virtual Symposium on Careers in Data Science and Statistics is scheduled for Tuesday, March 23, 5:00-6:00 pm. Lissa Callahan from ADM Associates will be the speaker; she will be joined by ADM colleagues for the discussion panel. The topic will be **Data Analytics in Energy**. Φ

K-12 Poster Competition

NV-ASA has regrettably found it necessary to cancel its K-12 Statistics Poster Competition due to the pandemic this year. In the past posters have been submitted physically and judged on the State level, with prizes awarded in four grade categories: K-3, 4-6, 7-9, and 10-12. Winners and honorable mentions in each category were forwarded to the national competition. NV entrants have very often received prizes and honorable mentions in the national competition.

The national competition will take place this year, but virtually. Posters will be submitted electronically, and must be received by April 1. For more information, please visit the Competition website <https://www.amstat.org/asa/education/ASA-Statistics-Poster-Competition-for-Grades-K-12.aspx>. 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

Job Opening at Wells Fargo

The job posting aims to fill a position on the Natural Language Processing (NLP) team within the Artificial Intelligence Center of Excellence at Wells Fargo. The role will involve partnering with audit leaders to design, develop, and deploy Natural Language and Machine Learning models. Candidates should have strong quantitative skills, programming experience (Python), knowledge of NLP, and ability to wrangle data. For more information and to apply, visit <https://www.wellsfargojobs.com/job/tempe/data-science-audit-model-development-quantitative-analytics-spec-3/1251/18639373>. %

Online M.S. in Business Analytics Program at UNR

During the Data Science Education Opportunities in Nevada portion of our Fall Virtual Symposium in 2020 we heard from Kal Joshi, describing the new Online M.S. in Business Analytics Program at UNR. A brochure and enrollment information are now available at www.onlinedegrees.unr.edu/msba. Σ

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. \star

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 (2020-2022)
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.

© 2021 Nevada Chapter of American Statistical Association

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



When Big Data is so Messy, You Never Get to Analyze It

Christopher Tong

The late biostatistician John C. Bailar once wrote, "My experience with data analysis indicates to me that problems with the data themselves are usually more numerous and more serious than problems with methodology." Indeed, when data problems are insurmountable, the most advanced analytics in the world may not get very far. Consider, for example, some observations by Arvind Krishna, currently CEO of IBM. (Incidentally, IBM's founder, Herman Hollerith, was a member of the American Statistical Association.) At the *Wall Street Journal's* 2019 "Future of Everything Festival", then IBM vice-president Krishna explained that "80% of the work with an artificial intelligence project is collecting and preparing data. Some companies aren't prepared for the cost and work associated with that going in." He continued "And so you run out of patience along the way, because you spend your first year just collecting and cleansing the data. And you say: 'Hey, wait a moment, where's the AI? I'm not getting the benefit.' And you kind of bail out." Naturally, he did not provide specific examples, but *WSJ* reporter Jared Council noted the cancellation of a major IBM Watson contract with MD Anderson Cancer Center in 2016, after the latter spent \$62 million on the project. Eliza Strickland published an extensive post-mortem of that episode in *IEEE Spectrum*, noting that "At MD Anderson, researchers put Watson to work on leukemia patients' health records—and quickly discovered how tough those records were to work with. Yes, Watson had phenomenal NLP [natural language processing] skills. But in these records, data might be missing, written down in an ambiguous way, or out of chronological order."

What prompts me to write about this now is the termination (in February this year) of Haven Healthcare, a joint venture created in 2018 by Amazon, JPMorgan Chase, and Berkshire Hathaway. Their goal was to revolutionize employee healthcare delivery for these three major US companies. With the managerial talent, financial expertise, and technological firepower of the sponsoring CEOs Jeff Bezos, Jamie Dimon, and Warren Buffett and their organizations, what could go wrong? Plenty, it turns out. Sebastian Herrera and David Benoit of the *WSJ* have their own post-mortem of this "epic fail", but I want to focus on one particular aspect they describe. "Data was a central challenge. Haven struggled to aggregate and analyze information on health-care costs for the three companies' employees. Data concerns from the partners and resistance from insurers stymied Haven's efforts to determine how much the companies paid for medical care and why." They continue, "Initially, the leaders of the joint venture imagined that if they could see what the three companies were spending on health care and why, the data would show them what to fix. ... Getting a hold of those figures proved difficult." The software platform Haven developed to pull this data together was rejected by the sponsors, and had to be rebuilt, "further delaying the goal of understanding, analyzing and reducing costs." While this was not the only reason for Haven's termination, it seems likely that the joint venture experienced what IBM's Arvind Krishna was warning about.

Riffing on R. A. Fisher's famous three problems of model specification, parameter estimation, and sampling distributions, Colin Mallows proposed the "Zeroth Problem": "Considering the relevance of the observed data, and other data that might be observed, to the substantive problem". I now propose a corollary to the Zeroth Problem: "If the available data are indeed relevant to the problem, how long will it take (and how costly will it be) to assemble, clean, and format the data into a structure amenable for analysis, if this is even possible?"

Sources

John C. Bailar, III, "Bailar's Laws of Data Analysis". *Clinical Pharmacology and Therapeutics*, 20:113-119 (1976).
Jared Council, "Data Challenges Are Halting AI Projects, IBM Executive Says". *Wall Street Journal*, 28 May 2019.
Sebastian Herrera and David Benoit, "Why the Amazon, JPMorgan, Berkshire Venture Collapsed: 'Health Care Was Too Big a Problem'". *Wall Street Journal*, 7 Jan 2021.
Colin Mallows, "The Zeroth Problem". *The American Statistician*, 52:1-9 (1998).
Eliza Strickland, "IBM Watson, Heal Thyself". *IEEE Spectrum*, April 2019. Published online as "How IBM Watson Overpromised and Underdelivered on AI Health Care."

StatsChats: Open-door research support at UNR

Paul Hurtado, Department of Mathematics & Statistics, UNR

In early 2016, another assistant professor, Kevin Shoemaker (Natural Resources & Environmental Science) and I wanted a better way to efficiently help the various grad students who were approaching us for help with their statistical analyses. Our goals quickly became clear: help students find useful answers to their stats questions, minimize time spent in meetings and office visits, and open these help sessions to other students to maximize information sharing.

Our solution was to host weekly open-door support sessions now known as "StatsChats", and we were joined by other interested faculty including Jessi Brown, Ken Nussear, Matt Forister, and Perry Williams. Below is an overview of "lessons learned" during the past few years. We hope that anyone looking to start a similar discussion group at their institution finds them helpful.

What is StatsChats? It is a weekly, 1-hour discussion group open to students and the rest of the campus research community. Questions at all levels are welcomed. Participants are encouraged to join us not just when they need help, but also to provide help to others or just be "a fly on the wall" there to learn. You drop in for an hour, give and receive whatever help you can, and then everyone heads their separate ways, no strings attached.

How is it organized? Our primary way of organizing StatsChats is an email list used to announce meeting dates and times. Participants also sometimes use it to give the group a heads-up if they plan to bring a problem to that week's meeting. We also have a webpage to give people basic information, but the email list is the primary mode of communication. Each semester we book a room on campus that is equipped with a computer, projector, and white board.

What are the meetings like? The great majority of our meetings involve a grad student, or other researcher on campus, starting us off with an informal overview of their problem. This is followed by a group discussion. If nobody brings a question or problem to the group, we will often chat briefly about future plans for the group and end early. Once or twice in a semester, we will also have a scheduled mini-tutorial of interest to the regular participants. Topics vary, and range from a technical discussion about particular methods of analysis to learning software like R. Not only are these tutorials useful learning opportunities, but they help familiarize people with StatsChats so that they are familiar with the resource and how to access it.

What factors contribute to the success of these discussions? First, "Location, location, location!". If our meeting room is in the same building as a lot of grad student offices, or in a familiar meeting or classroom space, we get a good turnout. If we're one or two buildings away in an unfamiliar room, turnout is noticeably reduced.

Second, having a consistent weekly schedule seems to help everyone integrate StatsChats into their work routine, but can also exclude those with conflicts like departmental seminars or joint lab meetings. We actively solicit scheduling feedback each semester to minimize that problem.

In Spring of 2020, the COVID-19 pandemic pushed everything campus-related online, and in Fall 2020 we took a rare break from hosting StatsChats. However, this spring, we have decided to run StatsChats online and alternate between two different meeting times to minimize schedule conflicts. We hope that this more flexible schedule (and two email reminders per week!) will offset some of the downsides of holding these discussions over zoom.

Third, we often have three or more experienced faculty, postdocs, and/or senior graduate students in the room. Participants get advice from more than just one person. We have found that the most productive discussions draw from the collective expertise and experience of the whole group, and that the graduate students are very capable and supportive of one another. Facilitating peer-to-peer interactions benefits everyone involved.

Lastly, and perhaps most importantly, we try to cultivate a laid-back, friendly and supportive atmosphere that encourages students to feel safe sharing their statistics-related struggles. We don't treat StatsChats like a formal consulting service, or a seminar, or a place to forge new collaborations. It's a space to get help, to help others, and to learn.