Fall 2021 For Members and Friends of the NV ASA

SILVER STATE-ISTICS

Nevada Chapter of American Statistical Association

Nevada Chapter Newsletter Vol. 18 No. 3

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 EVENT

We're looking forward to our Fall Symposium and Annual Meeting on October 16. See the next page for details. %

The President's Corner

Christopher Tong

Since our last newsletter, we have hosted two major virtual events: our traveling course on data science, machine learning, and deep learning; and a symposium on data privacy and confidentiality. Both were advertised nationally and were very well-attended. I offer abundant thanks to the speakers, as well as the volunteers who helped organize these events and kept them running smoothly. We have one more major event to come shortly: our fall symposium and annual meeting. We hope you will join us for our flagship annual event.

Our chapter's pandemic-era activities were recognized in the August issue of *Amstat News*, in the article "How to succeed now and in the post-pandemic future" by the ASA Council of Chapters Governing Board. They cited our activities twice in their suggestions of "successful activities" for chapters to consider pursuing. We've also submitted an article to ASA's *Chapter Chatter* about our pandemic-era activities.

Behind the scenes there has been even more activity. The Governance Committee has been working on updating the chapter constitution and reforming other chapter policies and procedures. Meanwhile our Treasurer is testing a new software for membership management, Wild Apricot. We are using this to register attendees for the annual meeting. Please be patient with us as we figure out how to use the system.

Finally, our chapter election is in progress. Please take the time to vote. It has been my fate to serve you as an appointed president in 2020, and an elected one in 2021, during the first two years (!) of the SARS-CoV-2 pandemic. I look forward to turning over the chapter presidency to new leadership very soon. Thank you for reading my comments in this newsletter over the past two years. Σ

Elections

By now you should have received a ballot for our annual elections if you are a member**. If you believe that you should have received one but have not, please contact the chapter immediately**. ∞

Fall Symposium

Our Fall Symposium will feature several invited speaker sessions. The first will be on Data Science careers. Invited speakers will be Will Thompson and Sam Havens, experts in natural language processing as part of a career path in data science. We will also have an update on data science education in NV.

The next session will discuss cross-pollination among data professions. Daniel Wright, Kimberly Jennison, and Kelley Wheeler will focus on psychometrics, and Juli Petereit will focus on Bioinformatics. Our president Chris Tong will be a discussant.

Next there will be a round table on software for teaching statistics and data science, featuring Julia Anderson, Sharang Chaudhry, and Chad Cross, with open discussion on this issue.

We anticipate student presentations as well. The day will end as usual with our Annual Business Meeting. ≤

Short Course on Data Science

NV-ASA was pleased to host the virtual short course **Introduction to Data Science, Machine Learning and Deep Learning (in R and Python)** July 23-24, 2021, with support from the ASA Council of Chapters. The instructors were Hui Lin of Google and Ming Li of Amazon. Dr. Li is currently a senior research scientist at Amazon and adjunct instructor at the University of Washington and has been active in the Quality and Productivity Section of the ASA. Dr. Lin is currently a quant researcher at Google and has held a variety of roles in data science and analytics. You can find the course materials at https://course2021.scientistcafe.com/course-syllabus/.

The two half-day course focused on popular machine learning and deep learning models related to artificial intelligence. Dr. Li started by discussing the many concepts that the term "Data Science" can be applied to, starting with "the discipline of making data useful", and involving the three tracks of engineering (data storage, management, and production), analysis (relating data to the real-world domain of interest and doing exploratory analysis and "story-telling", and modeling/inference (how do the data relate to the problem at hand, and how should we advise the stakeholder). After presenting some background on big data and cloud-based data, he walked us through some exercises using Databricks Community Edition (freeware) along with R and Python. The final topic for the first day was tree-based models, which encompasses a variety of things including but not limited to classification and regression, again with some hands-on experience with R and Python.

The second day featured neural networks, both convolutional and recurrent. Applications presented included visual object identification and classification via convolutional approaches to large two- and three-dimensional pixel data. Recurrent neural networks expand this methodology to real-time applications such as speech recognition. ΣΣ

Data Privacy and Confidentiality

This event took place Saturday September 11.

The first speaker was **Darren Toh** of the Dept of Population Medicine at Harvard Medical School and Harvard Pilgrim Health Care Institute. His title was **Protecting data privacy in multi-center studies: Experience from an Epidemiologist.**

We next heard from Thomas Kent, who is currently a federal government statistician, after exploring numerous positions in the US and Italy. His title was **An R package to implement secure multiparty distributed regression on vertically partitioned data**. One feature that he described is having a trusted third-party analysis center which handles communications between and codes used by data centers to implement various regressions. Their R package, called VDRA, is available on Github and CRAN.

Aleksandra Slavkovic̀ is in the Eberly College of Science at the Penn State University. Her title was **Valid statistical inference with privacy constraints**. The methodology she presented is called **differential privacy**, involving balancing the trade-off between

statistical inference quality and data privacy through various means, sometimes involving introducing some randomness (synthetic data).

Jordan Awan, a former student of Dr. Slavković and now at Purdue University, continues investigations into differential privacy. His talk was on **Canonical noise distributions and private hypothesis tests**, focusing on differences in proportions for binary data, and coming up with approaches providing nearly as much statistical power as procedures which disregard privacy issues and are in fact more powerful than classical normal approximations. λ

NV-ASA Chapter Governance and Constitutional Committee

The national Council of Chapters has recommended that all chapters review governance documents including chapter constitutions, and where they exist, bylaws and/or policies and procedures. In response, the NV-ASA ad hoc Governance Committee was created earlier this year to review the chapter's constitution. Factors such as the widespread use of the Internet, ongoing evolution in the field of statistics, and programs and initiatives fostered by ASA require limited amendment of verbiage in the 20-year old NV-ASA constitution. Per requirements, any proposed amendment in a chapter's constitution must be published prior to a chapter-wide member vote. To that end, the Governance Committee will introduce proposed amendment changes during the October Annual Meeting and all recommended changes will be published on the Chapter website. A vote on the proposed amendments will be held electronically sometime after January 1, 2022. Φ

Wild Apricot

Our chapter membership has increased considerably during the past couple of years. In order to better keep track of new members, membership renewals and calculating proper event fees based on membership categories, NV-ASA is engaged in a trial run of the **Wild Apricot** membership management software. Registrations for our Fall Symposium and new and renewing memberships are now being handled by our Treasurer using Wild Apricot. Ω

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

UNLV-UNR Statistics and Data Science Seminars

The UNR Dept of Math & Stat and the UNLV Dept of Math Sciences are holding bi-weekly virtual seminars on various topics in statistics and data science Fridays at 11:00. For details contact ania@unr.edu or kaushik.ghosh@unlv.edu. δ

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 (free 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 $200 ($100 for members of the national ASA).

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**. Otherwise, whether a national ASA member or not, you can join through our Wild Apricot site: **https://ncotasa.wildapricot.org/join-us**. Another option is to join while registering for an event such our Fall Symposium and Annual Meeting.

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

Floating Point Arithmetic and Data Science

Christopher Tong

Many statisticians and data scientists take floating point arithmetic for granted. We program our computers to carry out calculations and accept the results at face value. We rarely worry about the nuances of binary representation of numbers and their conversion to decimal form. Our attitude resembles that of the great 19th century electrical engineer and applied mathematician, Oliver Heaviside, who asked "Shall I refuse my dinner because I do not fully understand the process of digestion?" Well, did you know that floating point arithmetic is *not associative* (Demmel & Riedy, 2021)? Did you know that the oldest known trigonometric table, a 3700-year old clay tablet found in southern Iraq, is also allegedly the *most accurate* trig table known to exist? This is because Babylonian arithmetic is base-60, which can represent many fractions *exactly*, far more than in our base-10 system, in which many fractions must be approximated or rounded, like 1/3 = 0.33333.... (Mansfield & Wildberger, 2017).

Mishandling floating point numbers can result in catastrophic failure. Mathematician Douglas N. Arnold (1998) has documented the following two examples. The explosion of an unmanned Ariane 5 rocket in 1996 was traced to an overflow when a 64-bit floating point number was converted to a 16-bit signed integer. More tragically, 28 U.S. soldiers were killed in 1991 during the First Gulf War, when a Patriot missile in Saudi Arabia failed to intercept an Iraqi Scud missile. The error was traced to a loss of precision of time when tracking the Scud's velocity, as represented in the system clock's 24-bit register. Timing errors are particularly insidious. Faux and Godolphin (2021b) claim that rounding errors in quartz crystal stopwatches are theoretically capable of reversing the ranking of elite track or swimming athletes in a close race.

What about statistical calculations? In a 2011 stackexchange post, user "whuber" stated that the smallest *p*-value that can be represented in IEEE 754 double precision arithmetic is 10^-303. [A p-value that small might be considered a bit meaningless, because such a value is way smaller than $1/n$, where n = the number of observable particles in the universe (10^80 ish).] More practically, versions of Microsoft Excel prior to 2003 had a notorious flaw when calculating the sample variance (and therefore standard deviation). As we all know, there are two convenient, mathematically equivalent ways of expressing the variance:

$$(n\text{-}1)\ \text{Variance}(x_i) = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (1/n)(\Sigma x_i)^2.$$

As Hans Pottel (2000) noted however, in floating point arithmetic, the first form is more numerically stable than the second. He stated that the second formula “has the property that it suffers from subtractive cancellation for data sets for which the mean...is large compared to the standard deviation” and that an error analysis of the second formula shows that it has about twice “the number of incorrect significant figures in the results” compared to the first formula. However, the old versions of Excel used the unstable second form, while Excel 2003 and later use the more reasonable first form. Pottel (2000) shows that it is extremely easy to demonstrate the floating point error in old Excel. He gives an example of varying data that (old) Excel claimed had a precisely zero standard deviation. The solution is to manually subtract the mean from all data prior to applying the standard deviation function. (Subtracting a constant has no mathematical effect on the variance, but apparently it has a substantial effect in floating point calculations, the way old Excel coded it.) Earlier this year, Microsoft (2021) admitted that the variance algorithm had been changed for later versions of Excel. They also claimed that differences in outputs from the two algorithms are “likely to occur only in extreme situations.”

A personal anecdote. My very first “statistical consulting” project involved analyzing accelerator mass spectrometry measurements of Calcium-41 isotope in a study of bone resorption in a group of post-menopausal women. The technology was capable of measuring extremely small concentrations of this isotope in urine. Theory predicted the form of the highly nonlinear function that such data (calcium-41 to calcium-40 ratios on the order of 10^-10) should follow over time (samples were collected over several weeks), but four fitting parameters had to be estimated

empirically. I imported the data into SAS and tried to fit the model using the Levenberg-Marquardt algorithm. However, it would not converge. My statistics professor was baffled. However, I decided to "change the units of measurement", ie, multiply all the *y*'s by 10^10 to make them of order unity. SAS was now able to converge and estimate the fitting parameters. (The estimated parameters could then be re-expressed in the original units if desired.) I bet the SAS implementation of the iterative algorithm checked for convergence by comparing the reduction in sum of squared error with a default tolerance value, and thus could not handle *y*'s that may have been even smaller than that tolerance. The key was to not just think like a statistician, but think like a computer scientist too.

Hough (2019) and Demmel & Riedy (2021) provide some discussion about the latest (2019) revision of the IEEE 754 standard for floating point arithmetic. The standard was first issued in 1985. The 2008 revision included an interesting bug: the standard definitions of the Min and Max functions were not associative in the presence of "Not-a-numbers" (NaNs). The 2019 revision fixes this bug. Hough (2019) discusses the history of the IEEE 754 standard, while Demmel & Riedy (2021) focuses on the 2019 update as well as providing several examples of "epic fails". Faux & Godolphin (2021a) provide further examples of floating point failures from physics, astronomy, and computational science. (After reading these examples, I am not convinced that the "extreme situations" Microsoft spoke of necessarily implies "uncommon situations".)

It is anticipated that future floating point standards will need to explicitly accommodate machine learning and related computational tasks. Hough (2019) observes that such tasks "obtain greater accuracy by processing more data faster than by computing with more precision -- rather different constraints from those for traditional scientific computing. ... There might be arithmetic standards dedicated to very specific application areas, rather than compromises intended to be suitable for a wide range of diverse applications." Demmel & Riedy (2021) note that machine learning applications "benefit from a wider exponent range to represent smaller probabilities, thus leading to formats like Google's bfloat16 (with 7(+1) bits of precision, eight bits of exponent, and one sign bit)."
I would applaud a more fit-for-purpose approach to floating point arithmetic standards, but this places the burden on us, as data scientists and statisticians, to specify how *we* want *our* floating point arithmetic to work. How many of us understand floating point arithmetic deeply enough to even begin to address that question!

Exercise for the reader

Try the following simple calculations in your favorite spreadsheet or statistical software. Let $p = 769933891$ (a randomly selected 10-digit prime number), a constant, and let $q = \{3, 5, 7\} \times 10^{-9}$. Let $k = p + q$. Calculate the standard deviation of q and then calculate the standard deviation of k. Mathematically, these standard deviations should be identical, but in floating point arithmetic, you may get something quite different! I surmise that the title of a recent book by Hinch (2020) applies not only to fluid dynamics, but to statistics and data science, and any other field that involves computing.

References

D. N. Arnold, 1998: Some disasters attributable to bad numerical computing. https://www-users.cse.umn.edu/~arnold/disasters/.
J. Demmel & J. Riedy, 2021: A new IEEE 754 Standard for floating-point arithmetic in an ever-changing world. *SIAM News*, 54 (6): 9-11. https://sinews.siam.org/Details-Page/a-new-ieee-754-standard-for-floating-point-arithmetic-in-an-ever-changing-world.
D. A. Faux and J. Godolphin, 2021a: The floating point: tales of the unexpected. *American Journal of Physics*, 89: 806-814. https://doi.org/10.1119/10.0003915.
D. A. Faux and J. Godolphin, 2021b: The floating point: rounding error in timing devices. *American Journal of Physics*, 89: 815-816. https://doi.org/10.1119/10.0003919.

E. J. Hinch, 2020: *Think Before You Compute: A Prelude to Computational Fluid Dynamics*. Cambridge University Press.
D. G. Hough, 2019: The IEEE Standard 754: one for the history books. *Computer*, 52 (12): 109-112. https://doi.org/10.1109/MC.2019.2926614.
D. F. Mansfield & N. J. Wildberger, 2017: Plimpton 322 is Babylonian exact sexagesimal trigonometry. *Historia Mathematica*, 44: 395-419. https://doi.org/10.1016/j.hm.2017.08.001 See also video: https://www.youtube.com/watch?v=i9-ZPGp1AJE&t=127s.
Microsoft, 2021: Differences between the Excel statistical functions STDEVPA and STDEVP. https://docs.microsoft.com/en-us/office/troubleshoot/excel/statistical-functions-differences.
H. Pottel, 2000: Statistical flaws in Excel. https://www.asq904.org/StatisticalFlawsInExcel.pdf
"whuber", 2011: Response to "Sanity check: how low can a p-value go?" https://stats.stackexchange.com/questions/11812/sanity-check-how-low-can-a-p-value-go.