KISS Newsletter

January 1, 2016

Korean International Statistical Society

Volume 6 Issue 1

Inside this issue:

President's Corner

Dear KISSers,

Happy 2016, Year the Monkey!
I wish all members another prosperous and fruitful year.

In 2015, one of the most significant achievements was the career development workshop held at Joint Statistical Meeting (JSM) in Seattle. It attracted many young KISSers and panelists provided wonderful and thoughtful advices throughout the workshop. I would like to thank Dr. Mimi Kim and the panelists for their contributions to the successful workshop. You can read more details about the workshop on page 4. KISS is planning another workshop for young members at the upcoming JSM (Chicago). I hope that many members can attend it. It will be really useful to your career.

Another exciting thing was the establishment of the KISS early career awards. It is established to recognize KISS members in their early careers with outstanding productivity and potential to make significant contributions to the field. Six awards were given during the KISS annual meeting at JSM. I would like to thank Drs Jae-Kwang Kim and Mi-Ok Kim for their excellent work. KISS plans to give the awards again this year. I hope that many members who meet the award criteria can apply for the awards.

Many KISS members continued to win various awards as you can find them in this newsletter. In particular, Drs Chul W. Ahn and Yoonkyung Lee were elected as the ASA Fellows. Dr. Jae-Kwang Kim received Gertrude M. Cox Award. Dr. Jong-Min was awarded for his outstanding contribution to the field by the Korean Statistical Society. I am so happy whenever I hear such exciting news. I hope that we can hear many more exciting news in 2016.

I believe that next few years will be quite exciting to watch out for new waves of big data and data science. Big data have been a hot topic during the last few years and it will continue to be so. I have read many interesting publications featuring theoretical and applied statistical considerations for big data. There is no doubt that we will see many more publications on this hot topic. In addition, many new degree programs in data science, from Bachelor to PhD, have been launched at major universities across the USA. At some places, statisticians lead them, while we are left out at other places. It is unclear how data science will find its place among statistics, computer science, and informatics. With the balanced training in big data, statistics and computation skills, the new curricula in data science programs may be more attractive to the industry employers. Anyway, I think that it will be an exciting and turbulent time for statisticians.

Finally, 2016 will be the last year of my term as the president. It has been quite challenging and exiting years and we have achieved many things since 2010. I would like to thank all members, officers and board members for their love, services and strong supports for KISS over the years. I hope that this continues in 2016 and beyond. We always love to hear your ideas and suggestions how KISS can serve its members better. So, please do not hesitate to send them to me!

Cheers!
Dongseok Choi

2015 KISS Annual Meeting at JSM

The 2015 KISS annual meeting was held at the Joint Statistical Meetings on the August 10th in Seattle.

1. Special opening remarks: Professor Xuming He (University of Michigan)
2. The Executive Director Mi-Ok Kim reported the last year's KISS activities on behalf of all officers (see below for more details)
3. President's Invited Talk: New ASA Fellows Drs Chul Ahn (University of Texas) and Yoonkyung Lee (Ohio State University)
4. KISS Career Development Award
5. Business items and announcements:
   - Membership update
   - JSM 2015: KISS sponsored 5 sessions
   - KISS JSM 2016 Program Chair: Dr. Yoonkyung Lee (Ohio State University), JSM invited session proposals for KISS sponsorship (8/31/2015)
   - KSS spring/fall meetings, invited sessions
6. *Communications for Statistical Applications and Methods*: Looking for proposals for special issues, Being under review by SCOPUS for indexing, Need more submissions
7. Dinner: around 60 participants

KISS Officers' Reports

- KISS Election: President-elect (Dr. Mimi Kim), New Board members (Drs. Jong-Hyeon Jeong, Moonjung Cho, Hyunsik James Lee)
- KISS workshop on career development and mentoring: 8/9/2015
- Executive Director: Oregon title registration renewal ($50), Oregon CT-12 report ($10), IRS tax report.
- Communications Director: Mailing lists and the KISS website (http://statkiss.org/) have been updated. A newsletter was published in January, 2015.
- Treasurer: Report membership paid and KISS Financial report as of 5/31/2015

2014 Carry forward:	$12,974.83
2015 Income:	$ 4,148.25
2015 Expense:	$ 567.55
2015 Balance:	$16,195.53

- Program Chair: KISS is the main sponsor of the following 5 sessions:

8/9/2015 (8:30 AM) Computational Methods for Big Data and Visualization Problems

8/10/2015 (8:30 AM) Recent Advances in Theory and Methods for Hypothesis Test, Sampling, and Dimension Reduction

8/10/2015 (10:30 AM) Rediscovering Non- or Semiparametric Approaches to Longitudinal Data Analysis

8/11/2015 (10:30 AM) New Horizons of Quantile Regression Analysis: Longitudinal and Recurrent Event Data

8/12/2015 (10:30 AM) Contributed Oral Poster Presentations

Awardees at JSM 2015

ASA Fellow, 2015:
Dr. Chul W. Ahn (U of Texas)
Dr. Yoonkyung Lee (Ohio State U)

Gertrude M. Cox Award:

Jae-Kwang Kim (Iowa State U)

Dinner at JSM

2015 KISS Career Development Workshop

Dr. Mimi Kim

Over 40 people participated in the first KISS Career Development Workshop which was held at the 2015 Joint Statistical Meetings in Seattle. The workshop began with a panel discussion with senior statisticians from academia, industry, and a policy research organization: **Dr. Chul Ahn**, Professor of Biostatistics at UT Southwestern Medical Center; **Dr. Daniel Heitjan**, Professor of Statistical Science at Southern Methodist University and Professor of Clinical Sciences at UT Southwestern Medical Center; **Dr. Donsig Jang**, Director of Data Science and Statistics at Mathematica Policy Research; and **Dr. Sunhee Kwon Ro**, Director of Biostatistics at Amgen, a leading biotechnology company.

The four distinguished panelists talked about the challenges they faced when they were starting out in their careers, how they overcame cultural and language barriers on the job, and how they developed their leadership skills. A recurring theme was the importance of persistence in achieving career goals and not being discouraged by failure; one panelist told the story of how Harvard Statistics Professor Donald Rubin initially received multiple rejections from top journals when he was trying to publish his seminal work on "*Inference and Missing Data*." It eventually appeared in Biometrika in 1976 and became one of his most widely cited papers.

During the workshop, the issue of the "bamboo ceiling" was also discussed, a phrase which was coined by Jane Hyun, a Korean-American executive coach, who wrote a book on the combination of individual, cultural, and organizational factors that can make it difficult for Asians to advance in their careers. Dr. Heitjan, who has mentored many Asian students, emphasized not only the importance of strong communication skills but also making the effort to be knowledgeable about American culture and current events to facilitate social interactions with non-Korean colleagues and to better assimilate into the work environment.

Dr. Xihong Lin, Chair of the Department of Biostatistics at Harvard Chan School of Public Health and recent recipient of the prestigious Outstanding Investigator Award from the National Cancer Institute, also spoke briefly about her professional experiences and offered several valuable career tips. She said that strategizing and focusing too narrowly on the final goal, e.g., tenure, is stressful and counterproductive; instead, one should find fulfillment from the work itself, which should be done with integrity, and positive results will follow. She encouraged junior statisticians to get involved in their department's activities and volunteer for positions, such as seminar organizer, which provide opportunities to get to know senior statisticians at other institutions. She also pointed out that it is easier to make an impact if you can find an emerging methodological area to work in rather than going into a crowded field. Any new approach in an established area, such as predictive modeling, will take time to develop since it would have to be compared to the numerous methods that are already available.

According to many participants, the most enjoyable aspect of the workshop was the opportunity after the panel discussion to interact with the speakers and other attendees in the small group breakout sessions and share professional experiences and concerns. Overall, the event was deemed to be a success and participants expressed enthusiasm for another workshop that could focus on practical issues such as how to find a job, pros and cons of different work environments, preparing for an interview, and negotiation skills. In addition to these career workshops, the KISS mentoring program which was initiated two years ago is still active and ongoing so anyone who is interested in participating either as a mentor or mentee should contact Dr. Mimi Kim.

2015 KISS Career Development Award

Dr. Jae-Kwang Kim

KISS Career Development Award was established in 2015 to recognize statisticians who are in the early stages of their careers and who have demonstrated outstanding productivity and the potential to make significant contributions to the field of statistics. Each awardee received $500 cash prize. The awardees are Sangbum Choi, Hangjoon Kim, Jaeun Choi, Seonjin Kim, Hyokyoung Hong, and Dongjun Chung.

Dr. Sangbum Choi is an assistant professor in Division of Clinical and Translational Science, the University of Texas at Houston. Sangbum got his PhD from University of Wisconsin at Madison in 2010 and did a three year postdoc work at the University of Texas MD Anderson Cancer Center before he joined the current position in 2013. His research area lies in semiparametric methods with missing data and survival analysis.

Dr. Hang Joon Kim is currently an assistant professor of Department of Mathematical Science at University of Cincinnati. Hang Joon got his PhD from Ohio State University in 2012 and then worked as a post doc at National Institute of Statistical Science and Duke University. His research area is Bayesian analysis and missing data imputation.

Dr. Jaeun Choi is currently an assistant professor of Division of Biostatistics, Department of Epidemiology and Population Health, in Albert Einstein College of Medicine. Jaeun got her PhD from University of North Carolina in 2011 and then worked as a post doc at Harvard Medical School. Her research area is in survival analysis and causal inference.

Dr. Seonjin Kim is currently an assistant professor of Statistics department at Miami University. Seonjin got his PhD from Penn State University in 2013. His research area is quantile regression and nonparametric method.

Dr. Hyokyoung Hong is currently an assistant professor in Department of Statistics and Probability at Michigan State University. Hyokyouong got her PhD from Univ. of Illinois at Urbana-Champaign in 2008 and then worked at City University of New York before she joined at her current position. Her research area is in quantile regression and variable selection.

Dr. Dongjun Chung is currently an assistant professor in Department of Public Health Sciences at Medical University of South Carolina. Dongjun got his PhD from University of Wisconsin at Madison in 2012 and then worked as a post doc at Yale University. His research area is statistical genomics and genetics.

2015 KISS Career Development Award Committee

Chair:
Jae-Kwang Kim
(Iowa State U)

Members:
Mi-Ok Kim
(Cincinnati Children's Hospital Medical Center)

Dongseok Choi
(Oregon Health & Science University)






ASA Statement on the Role of Statistics in Data Science

Statement Contributors:

David van Dyk, Imperial College (Chair)

Montse Fuentes, NCSU

Michael I. Jordan, U.C. Berkeley

Michael Newton, University of Wisconsin

Bonnie K. Ray, Pegged Software

Duncan Temple Lang, U.C. Davis

Hadley Wickham, RStudio

The rise of data science, including Big Data and data analysis, has recently attracted enormous attention in the popular press for its spectacular contributions in a wide range of scholarly disciplines and commercial endeavors. These successes are largely the fruit of the innovative and entrepreneurial spirit that characterize this burgeoning field. Nonetheless, its interdisciplinary nature means that a substantial collaborative effort is needed for it to realize its full potential for productivity and innovation. While there is not yet a consensus on what precisely constitutes data science, three professional communities, all within computer science and/or statistics, are emerging as foundational to data science: (i) Database Management enables transformation, conglomeration, and organization of data resources, (ii) Statistics and Machine Learning convert data into knowledge, and (iii) Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.

Certainly, data science intersects with numerous other disciplines and areas of research. Indeed, it is difficult to think of an area of science, industry, commerce, or government that is not in some way involved in the data revolution. But it is databases, statistics, and distributed systems that provide the core pipeline. At its most fundamental level, we view data science as a mutually beneficial collaboration among these three professional communities, complemented with significant interaction with numerous related disciplines. For data science to fully realize its potential requires maximum and multifaceted collaboration among these groups.

Statistics and machine learning play a central role in data science. Framing questions statistically allows us to leverage data resources to extract knowledge and obtain better answers. The central dogma of statistical inference, that there is a component of randomness in data, enables researchers to formulate questions in terms of underlying processes and to quantify uncertainty in their answers. A statistical framework allows researchers to distinguish between causation and correlation and thus to identify interventions that will cause changes in outcome. It also allows them to establish methods for prediction and estimation, to quantify their degree of certainty, and to do all of this using algorithms that exhibit predictable and reproducible behavior. In this way, statistical methods aim to focus attention on findings that can be reproduced by other researchers with different data resources. Simply put, statistical methods allow researcher to accumulate knowledge.

For statisticians to help meet the considerable challenges faced by data scientists requires a sustained and substantial collaborative effort with researchers with expertise in data organization and in the flow and distribution of computation. Statisticians must engage them, learn from them, teach them, and work with them. Engagement must occur at all levels: with individuals, groups of researchers, academic departments, and the profession as a whole. New problem-solving strategies are needed to develop “soup to nuts” pipelines that start with managing raw data and end with user-friendly efficient implementations of principled statistical methods and the communication of substantive results. Statistical education and training must continue to evolve—the next generation of statistical professionals needs a broader skill set and must be more able to engage with database and distributed systems experts. While capacity is increasing within existing and innovative new degree programs, more is needed to meet the massive expected demand. The next generation must include more researchers with skills that cross the traditional boundaries of statistics, data-

bases, and distributed systems; there will be an ever-increasing demand for such "multi-lingual" experts.

Working with statisticians, departments of statistics, and other professional societies, the American Statistical Association (ASA) is well positioned to help formulate discussion around the role of statistics in data science, to navigate the way forward in this quickly evolving environment, and to provide forums of communication and collaboration among data scientists, including statisticians and nonstatisticians alike. The ASA aims to facilitate collaboration between statisticians and other data scientists and thus enable them to achieve more than they could on their own.

* Many universities have launched data science programs. A few examples are introduced in the AMSTAT magazine: http://magazine.amstat.org/blog/2015/08/01/new-undergraduate-data-science-programs-2/

The following article introduces the program in the Oregon State University.

The second article on data science was written by Debbie Farris, and Sastry Pantula contributed to this article.

* KISS obtained permission to reproduce this article from the Oregon State University.

Data Science: Extracting knowledge from a torrent of information

Everything from health records, environmental monitoring, agriculture and online behavior with clicks, "likes," and purchases generate data every second. With this proliferation of data, the ability to analyze large data sets—big data– has become a platform of competition. It is a key driver of productivity, innovation and market demand. A panel of the United Nations Secretary General recently reported that for too long global development efforts have been hampered by a lack of the most basic data about the social and economic circumstances in which people live. Data science is a key area of growth and investment for the College of Science and for Oregon State because it is highly relevant, we have an obligation, we have key strengths and there is tremendous opportunity.

First, big data is highly relevant in a 21st century world. It satisfies a growing need to manage, analyze and interpret massive, complex data sets to solve problems and to better inform decision makers across disciplines. Because data analysis techniques are complex, the meaning can be misunderstood by those charged with prioritizing, designing and leading public policy.

Angus Deaton, the 2015 Nobel Laureate in Economics, spoke recently about the importance for better data that leads to better lives. Understanding patterns in large data sets is extremely important and has tremendous impacts on our world. OSU's Strategic Plan 3.0 outlines its commitment to leveraging technology as a strategic asset:

"Technology and information occupy a critical role in a 21st century university… Greater accountability, enhanced expectations of a current generation and growth in the development, management and delivery of digital resources point to the expanding role that big data, analytics and information technologies provide as a strategic enabling asset."

By aligning with national and global priories for big data, the College of science is able to lead big data analytics at OSU and beyond.

"Data science is the heartbeat of 21st century global economies, and innovations in sciences, engineering, business, and education are becoming increasingly computationally– and data-enabled." explains Sastry G. Pantula, dean of the College of Science. "Strategic investments in data analytics research and in training future data scientists will have long-term payoffs not only for our students, but also for industry and society."

Secondly, we also have an obligation and a responsibility to educate the next generation of data scientists with computational-thinking and data analytics skills to solve our most pressing challenges as part of a 21st century workforce. In a landmark report on Big Data, McKinsey & Company forecasted a national shortfall of 150,000 master's level professionals trained in data analytics with the ability to manage big data. To address this shortage, the College is developing undergraduate courses and is creating an online master's program in data analytics.

Third, data science capitalizes on our strengths and hallmark collaboration while transcending disciplines, moving seamlessly between research and classrooms. Data science will expand the university's footprint, positioning it as a leader in the statistical, mathematical and computational sciences. The College is developing a distinct research and education program in data sciences that integrates OSU strengths in computer science, genomics, statistics, mathematics, and applied sciences and policy.

Strategic investments in mathematics, statistics and life sciences faculty have extended the College's impact of data science on transdisciplinary research. In a science-without-borders approach, the College is deepening engagement between data science and other sciences, engineering, education, arts and business. Cluster hiring in bioinformatics across disciplines has brought expertise in mathematical biology; ecological, evolutionary, and functional properties of the microbiome; and deep sequencing data.

And finally, data science offers abundant opportunity. By aligning our expertise with market and national needs, federal priorities and funding opportunities, the College and OSU will advance the White House's National Big Data Research Initiative, which seeks to accelerate the pace of discovery in STEM and transform teaching and learning by improving our ability to extract knowledge and insights from large, complex collections of digital data.

Last year the federal government allocated $200 million for R&D in big data. Funding agencies are following suit. NSF has encouraged "research universities to develop interdisciplinary graduate programs to prepare the next generation of data scientists and engineers." The National Institutes of Health created Big Data to Knowledge, an initiative to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge.

In a boon to OSU's marine science and big data initiatives, NSF recently awarded OSU its NSF Research Traineeship award to build cohorts of leaders in marine science, data and policy. The five-year, $3 million award will prepare a new generation of natural resource scientists and managers who will combine mathematics, statistics, and computer science with environmental and social sciences to study, protect and manage ocean systems.

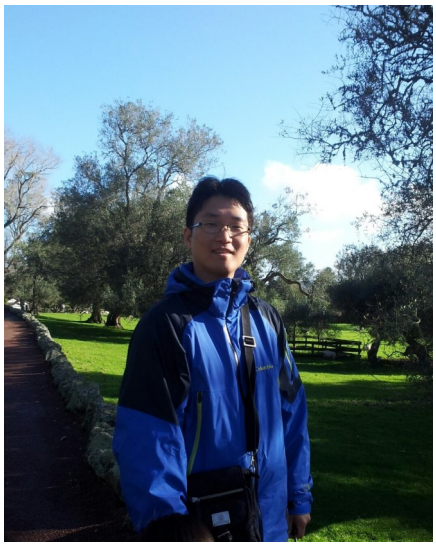
KOREAN INTERNATIONAL STATISTICAL SOCIETY

Member's Profile: Dr. Yoonsuh Jung

Before I moved to U.S., I was born, grew up, and got married in Seoul, Korea. I obtained B.A. from the Department of Statistics, Korea University. I completed my compulsory military service and retired as a sergeant in 2001. The military service was quite unusual experience as I have spent most of my life in schools. The second unusual life began in 2007 with the birth of my daughter. The first three years (2004 - 2007) in Columbus can be summarized as just studying Mathematics.

My main research areas are robust statistical modeling, quantile regression, high dimensional modeling, and variable selection (or model selection). My Ph.D. work was focused on methodology, and postdoc projects were both on methodology and its application to genetic data. In general, I am interested in developing statistical methodologies for better statistical modeling and decision procedure. Recently, I began to have more interest in the practical usefulness of my future research, and providing helpful solutions to the problems in the real world. Please visit my webpage for the details of research and professional activities.

(without fun). During the second half of my Ph.D., doing research became quite interesting, and I decided to find a job in academic area, which led me to do postdoc. Right after I moved to Houston for postdoctoral work, we had a boy in 2010. It seemed that having a family of four members soothed my homesickness. During postdoc period, I had a chance to work with several people including Prof. Jianhua Huang in Texas A&M University and Prof. Xuming He in University of Michigan. In this period, I felt I was turning to an independent researcher, and conducting research began to provide lots of fun and excitement. After I spent three years in Houston, I moved to New Zealand in June 2013. Among many advantages in the University, being treated fairly and reasonable amount of work loading have been the most attractive points. There are six faculty members in the Department of Statistic, but soon will be amalgamated with Department of



(www.stats.waikato.ac.nz/~yoonsuh)

Yoonsuh Jung (정윤서) is a lecturer (Assistant Professor equivalent) in the Department of Statistics at University of Waikato, New Zealand. In June 2010, he received his Ph.D. in Statistics at the Ohio State University, Columbus under the supervision of Prof. Steven MacEachern and Prof. Yoonkyung Lee. Between 2010 and 2013, he worked at MD Anderson Cancer Center as a postdoctoral fellow with Prof. Jianhua Hu.

Member's News

- Dr. Jong-Min Kim (U of Minnesota at Morris) received an award from KSS for his contribution to promoting statistical research.
- Eunjee Lee (U of North Carolina) and Sunghwan Kim (U of Pittsburgh) received the 2015 ENAR Distinguished Student Paper Award.
- Seyoung Park (U of Michigan) received the 2015 Statistical Learning and Data Mining Section Student Paper Award.

A Short Note on Large Sample Problem

Dr. Jai Won Choi

We discuss large sample when sample size is too big to handle with current method. As the data size increases rapidly in recent years, we use increased computing power to handle large samples. We may roughly divide all sizes of data into three sizes: small size (<20), middle size (20-300), and large size (>400). Here, the boundary of these divisions is very arbitrary as it also depends on the other factors such as population size N as well as precision, type of data, and other conditions. We have smaller problem for the middle size to use current methods (e.g. t-test or normal test) for statistical inference, but we have problem for other two, small and large, especially for large sample data, and discuss these two sizes.

For large sample (>400), as the current method depends on variance (except some nonparametric methods), and the variance in turn depends on sample size n. For large n >400, resulting variance is too small, and consequently test result is unnecessarily significant. So we randomly divide the large sample into groups of middle size and then apply traditional method to each group for testing. DeGroot and Schervish (2002) briefly mentioned several possible suggestions (p529-530) for large sample, One of suggestions is make the alpha level much smaller than traditional 0.01, 0.05 to fit the large sample. Another is replacing the single value of mu (mean) in the null hypothesis by an interval. Third is regard statistical problem as one of the estimation rather than one of testing hypothesis. However, these are not developed for practical application.

For a large sample or data set (>400), we may divide it into random groups of middle size (20-300) and then perform current testing on each group. If 95% (or 98% or another appropriate %) of the tests are significant, we may conclude the result is significant. If there are too many groups to handle (say millions), then we may use only a random sample of the groups. For small data (<20), test may be unreliable if the sample size is too small. For the small sample, we do not have enough information to make proper inference based on assumed distribution. We need more research in this area for statistical inference. One may use a distribution free method.

Small sample to Large sample

R.A. Fisher (1890-1962) studied Mathematics at Cambridge University from 1909. Among many accomplishments, he published the related papers to maximum likelihood estimation during 1912-1922, laying the foundation of current statistics. He developed the design of

experiment at Rothamsted Research center. He used the terms "variance" and "analysis of variance" for the first time. Fisher used small sets of data in his studies not exceeding more than he could calculate with his pen or pencil. For example, a few plots were used for his agricultural experimentations at Rothamsted. **Lady's tasting tea** is another good example. It is a randomized experiment reported in his book *The Design of Experiments* (1935). He used 8 cups, putting tea first in 4 cups and cream first in 4 cups randomly. He asked Ms. Muriel Bristol to identify which, tea or cream, was the first in his randomized blind test. The test used was Fisher's exact test. Here he used only 8 cups and calculated 70 combinations of 4 cups out of 8 cups. This could be the maximum numbers he could calculate at the time of no calculator or computer.

In recent years, data size has exploded and there is no way we can calculate with pen or even calculator. For example, Affordable health care needs to control the cost of the plan. To do so, they need to know the number of doctor visits. National Health Interview Survey (1990) used a sample of about 12,000 people from 300 million U.S. population and it reported that each person visited doctors 4 times a year excluding hospital inpatients. It is 48,000 visits in all. To complete cost calculation, need not only number of visits but also the related information such as whom and why they visited, and how much they paid for their visits and medications. The final data could be increased to millions. In this case, even if a small fraction of it, its data size will be prohibitively large. No current methods would work with such a huge sample size. Yet no question ever asked until recently (Choi, 2011) for testing hypothesis under the large sample situation.

An Example

Often the proportion from a clinical trial is used to prove the efficacy of a drug, we need the variance of this proportion p, in testing hypothesis P=0, which depends on n. If p is very small, we need a bigger n to detect the efficacy. But no matter what the size of proportion p is, (if n is greater than 300), test result is significant since these tests depend on the size of n. Conventional tests works only if the sample size is within this middle range at most. Table below shows standard deviation and test scores for n size 5,10, 20, 50, 100, 400, 500, and 1,000, and proportion p=0.1, 0.2, 0.4, 0.5, 0.6, 0.6, 0.8, and 0.9. (ss=$p(1-p)/n$, s=sqrt(ss), z=p/s when testing null hypo. H_0=0.)

Note that sample size n influences the test scores. For example, for p=0.1, n=5, we see z=0.745, for the same proportion p=0.1, z=2.108 for n=40, z=3.333 for n=100, z=6.666 for n=400, z=7.45356. Here, there is no other factor influencing the z scores except sample size n. At n=100, z=3.333 is significant at alpha= 0.025 even with this lower proportion 0.1. If sample size is greater than 300, the test result is always significant even for smaller

		p=0.1	p=0.2	p=0.4	p=0.5	p=0.6	p=0.8	p=0.9
n=5	S	0.13416	0.17889	0.21909	0.22361	0.21909	0.17889	0.13416
	Z	0.74536	1.11803	1.82574	2.23607	2.73861	4.47214	6.70820
n=10	S	0.09487	0.12649	0.15492	0.15811	0.15491	0.12659	0.09487
	Z	1.05409	1.58114	2.58199	3.16228	3.87298	6.32456	9.48686
N=20	S	0.06708	0.08944	0.10954	3.65148	0.10954	0.08944	0.06708
	Z	1.49071	2.23607	3.65148	4.47214	5.47723	8.94427	13.4164
N=40	S	0.04743	0.06325	0.07746	0.07906	0.07746	0.06325	0.04743
	Z	2.10819	3.16228	5.16398	6.32456	9.74597	12.6491	18.9737
n=100	S	0.03	0.04	0.04899	0.05	0.04899	0.04	0.03
	Z	3.33333	5.0000	8.16497	10.0000	12.2475	20	30
N=400	S	0.015	0.02	0.02449	0.025	0.02449	0.02	0.015
	Z	6.66667	10	16.3299	20	24.4949	40	60
N=500	S	0.01342	0.01789	0.02191	0.02236	0.02191	0.01789	0.01342
	Z	7.45356	11.1803	18.2574	22.3607	27.3851	44.7214	67.0820
N=1000	S	0.00949	0.01265	0.01549	0.01581	0.01549	0.01265	0.00949
	Z	10.5409	15.8115	25.8199	31.6228	38.7289	63.2456	94.8683

proportions. Therefore, we cannot use current testing for n is over 300.

Resampling hides real problems

National Health Interview Survey (NIHS, 1990) had the national sample of about 12,000 people and National Health and Nutrition Examination Survey (NHANES III) had 33,000 sample persons, encountered large sample problems. When the conventional tests did not work, some tried resampling methods: Balanced Half Sample (McCarthy, 1965), Jackknife, or Bootstrapping. As a practicing statistician, these resampling methods often used to calculate variances in NCHS. It only hides the large sample size problems. Either BHS or Jackknife does not produce correct variance for large sample situation. Bootstrapping is to take samples from an original sample. It does not give any better information than the original sample itself. These do not solve the large sample problem sample size is over million records.

We have unlimited computing power like cloud computing (for example **IBM Watson** Technology) www.**ibm**.com/outthink. **Watson** Is reinventing The Way We Work, Discover More, Data Intelligence, Cognitive Technology, Cognitive Innovation, and Watson Ecosystem. Even if the data size is over billions, the calculation is not a problem with currently available computing power.

Forming random groups

One Ph.D. student presented her research results. The sample sizes of her studies were over 1,500 and her test results were all very significant. I suggested that she could form 15 random groups of 100 to see how many of the groups are still significant. If 95% of them (about 14 out of 15) are still significant, she may conclude it is actually significant. Here, the group size is 100, but it could be any size from 20 to 300 (say). Which size is better? Since the sample size n is over 30, the t-distribution become almost identical to the normal distribution, we may try to use size n=30. We will further discuss it in another place.

We can randomly divide NHNES III sample of 33,000 into **110** groups of n=300, **165** groups of n=200, **330** groups of n=100, **660** groups of n=50, **1650** groups of n=20. Of course, the n size depends on several factors, for example, the rarity of cases, variation, significance of test result you want. Apply current test (t-test or normal test) to each group of n=20, 50, 100, 200 and 300. Since sample variances depend on sample size, as seen above, the bigger sample size gives more significant test results than the smaller sample does. Large % of significant tests will be resulted for large n size than smaller n. Because the variance of large sample shrinks too much just like in the table above, consequently, the testing result becomes significant unnecessarily.

One may use a new statistics instead, not depending on the variance (or n size). A type of nonparametric approach can be considered replacing traditional testing. I throw this to the readers hoping that one can come up with better solution for the large sample problem.

References

Choi, Jai Won (2011), A Thought on the Current Statistics. News Letter of Korean Statistical Association. April, 2011. p23-26

DeGroot, M.H. and Schervish, M.J. (2002), Probability and Statistics (3rd ed), Addison Wesley

Efron, B. and Gong, G (1983), A leisurely look at the bootstrap, the jackknife and cross-validation. Am. Statistician 37:36-48, 1983.

Fisher, R.A. (1937) The Design of Experiments (2 ed.). Edinburgh: Oliver and Boyd. 1937

McCarthy, P.J. (1982), Estimated Variance for the Combined Ratio Estimate Stratified, Two~stage Samples Without Replacement. B. V. Sukhatrne Memorial Lecture. Iowa State University, Apr. 16, 1982.

Prinz, F., Schlange, T., and Asadullah, K. (2011) Believe it or not: how much can we rely on published data on potential drug targets?. Nature Reviews Drug Discovery, 10, 712

SurveyMonkey—Official Website—SurveyMonkey.com

Call for Papers: Communications for Statistical Applications and Methods

Dr. Chul H. Ahn

Communications for Statistical Applications and Methods is an official journal of the Korean Statistical Society and Korean International Statistical Society beginning in 2013. Abbreviated title is 'CSAM'. It contains original articles dedicated to research in various fields of statistics and probability, or contributing to applied statistics through innovative data analysis and interpretation. Articles dealing with statistical education and software as well as survey papers are also welcomed. The journal welcomes articles from all countries. Our objective is to increase the visibility of CSAM journal by growing its content and distribution. By continually raising the quality of the journal and thereby increasing the likelihood of citation, we are working hard to list this journal in Science Citation Index Expanded (SCIE). We encourage all KSS and KISS members to submit papers via CSAM to reach this goal in the near future.

The journal accepts articles written in English and is published bi-monthly in January, March, May, July, September, and November. All of the manuscripts are peer-reviewed and all published articles are also freely available online. CSAM welcomes only original research articles for the form of publication. To submit your paper, please visit http://csam.or.kr

Upcoming Meetings

- The 14th Asia Pacific Bioinformatics Conference January 11-13, 2016, San Francisco Bay Area, CA, USA http://www.sfasa.org/apbc2016/apbc2016.html
- ENAR 2016, March 6-9, 2016, Austin, TX, USA, http://www.enar.org/meetings/spring2016/index.cfm
- KSS 2016 Spring Meeting (5/27-28, Kyungpook National University) Daegu, Korea, http://www.kss.or.kr/ (Note: there will be a KSS invitation session.)
- IBC/WNAR 2016, July 10-15, 2016, Victoria, Canada, http://biometricconference.org/
- 2015 5th IMS FIPS Workshop, June 25-27, 2015, Rutgers University, NJ, USA http://www.fsrm.rutgers.edu/fips2015
- The 4th Institute of Mathematical Statistics Asia Pacific Rim Meeting, June 27-30, 2016, Hong Kong, China https://ims-aprm2016.sta.cuhk.edu.hk/
- Joint Statistical Meetings 2016, July 31- August 4, 2016, Chicago, IL, USA, https://www.amstat.org/meetings/jsm/2016/
- The 2nd Pacific Rim Statistics Conference in Production Engineering Data, December 15-16, 2016, Seoul National University, Seoul, Korea
- The 10th ICSA International Conference on Global Growth of Modern Statistics in the 21st Century, December 19-22, 2016, Shanghai Jiao Tong University, Shanghai, China, http://www.math.sjtu.edu.cn/conference/2016icsa/Default.aspx

Korean International Statistical Society

0841 SW Gaines St. Unit 502
Portland, OR 97239
E-mail: info@statkiss.org
http://www.statkiss.org

If you would like to write an article or have comments for KISS newsletters, please email to info@statkiss.org.

Become a KISSer!

KISS has launched the web-site at **http://statkiss.org**. It contains the information about KISS, News and events, job opportunities, upcoming and past meetings, membership online registration, guest board, and photo gallery. Here is a brief membership online registration Instruction:

1. Please review different membership types from under Membership in the KISS website: http://statkiss.org/Membership.php

2. Please fill out the membership form under Membership/Form: http://statkiss.org/Form.php
'*' represents required fields

3. If you click "Submit", the screen shows the following message: "Thanks for submitting the form. If you would like to pay your membership fee by credit card online, please click here."

4. If you are not a student and would like to pay the membership fee online, please click "please click here". Then it will lead you to Membership/Payment: http://statkiss.org/Payment.php

5. Please select one membership type and click "Submit", then you will be forwarded to PayPal website.

6. You can pay as a guest (if you do not have an account with PayPal and do not want to create a new one), or you can login with your PayPal account to pay.

We believe that you may get an error when you try to pay with a credit card as a guest, but the credit card was registered for a PayPal user: e.g.. Your spouse is a PayPal member and registered a credit card, and you tried to pay with the same card as a guest.

KISS Korean International Statistical Society