Statistics Digest

THE NEWSLETTER OF THE ASQ STATISTICS DIVISION

VOL. 39, NO. 2 | JUNE 2020

asq.org/statistics

IN THIS ISSUE

Articles

COLUMNS

Departments

Message from the Chair

Amy Ruiz

It goes without saying that this year has had its fair share of unprecedented hardships and challenges. But as with any tribulation, there are lessons to be learned, opportunities to seize, and growth to achieve. It seems we are beginning to see the light shining at the end of the tunnel and we are confident that we will come out better on the other side.

With that said, the Stat Division has faced our own series of obstacles this year, but we are looking for opportunities in spite of these challenges. While WCQI was cancelled, a virtual 'conference' was held and was a success with over 7,453 ASQ members registered and 597 non-members registered. 4,465 members participated and the overall satisfaction rating (4s and 5s) was 82%. While FTC is cancelled for 2020, we can use the extra time this year to plan 2021 meetings and conferences and aim at making FTC 2021 the best one yet! The ASQ Canada Conference, which was set for October 19–20 this year will be cancelled as well, but there may be a virtual conference held with various speakers over a span of a few days. Be sure to check our website (links below) for additional details.

Additionally, we have had great success with our webinars this year. In January, Norma Antunano presented a Spanish webinar on 'Comparing Methods to Represent and Analyze Data' and in March, Daksha Chokshi presented a webinar titled 'Statistical Process Control: Myths, Misconceptions and Applications' to a virtual audience of 313. We have a few other exciting webinars planned this year so stay tuned!

I would like to extend my sincerest thank you to all of our members and well as our entire leadership team for not just `hanging in there', but for going the extra mile to ensure an innovative and productive year in spite of the many 2020 hurdles. Our leadership volunteers truly are the backbone of the Statistics Division and I am very grateful to each of them for their hard work and dedication.

Also, if you have not done so already, be sure to join myASQ and follow the myASQ Statistics Division Community at https://my.asq.org/communities/home/177. As a valued member of our online community, you will have access to many benefits and resources including:

- Events including Division conferences, Webinars, and related information at https://my.asq.org/communities/events/177
- Technical resources including the Statistics Digest, YouTube channel video series, and more at https://my.asq.org/communities/reviews/177

Continued on page 3

Statistics Division
ASQ Excellence Through Quality™

Submission Guidelines

Mini-Paper

Interesting topics pertaining to the field of statistics; should be understandable by non-statisticians with some statistical knowledge. Length: 1,500-4,000 words.

Feature

Focus should be on a statistical concept; can either be of a practical nature or a topic that would be of interest to practitioners who apply statistics. Length: 1,000-3,000 words.

General Information

Authors should have a conceptual understanding of the topic and should be willing to answer questions relating to the article through the newsletter. Authors do not have to be members of the Statistics Division. Submissions may be made at any time to newsletter@asqstatdiv.org.

All articles will be reviewed. The editor reserves discretionary right in determination of which articles are published. Submissions should not be overly controversial. Confirmation of receipt will be provided within one week of receipt of the email. Authors will receive feedback within two months. Acceptance of articles does not imply any agreement that a given article will be published.

Disclaimer

The technical content of material published in the ASQ Statistics Division Newsletter may not have been refereed to the same extent as the rigorous refereeing that is undergone for publication in *Technometrics* or *J.Q.T.* The objective of this newsletter is to be a forum for new ideas and to be open to differing points of view. The editor will strive to review all articles and to ask other statistics professionals to provide reviews of all content of this newsletter. We encourage readers with differing points of view to write to the editor and an opportunity to present their views via a letter to the editor. The views expressed in material published in this newsletter represents the views of the author of the material, and may or may not represent the official views of the Statistics Division of ASQ.

The Statistics Division was formed in 1979 and today it consists of both statisticians and others who practice statistics as part of their profession. The division has a rich history, with many thought leaders in the field contributing their time to develop materials, serve as members of the leadership council, or both. Would you like to be a part of the Statistics Divisions' continuing history? Feel free to contact chair @asqstatdiv.org for information or to see what opportunities are available. No statistical knowledge is required, but a passion for statistics is expected.

Vision

The ASQ Statistics Division promotes innovation and excellence in the application and evolution of statistics to improve quality and performance.

Mission

The ASQ Statistics Division supports members in fulfilling their professional needs and aspirations in the application of statistics and development of techniques to improve quality and performance.

Strategies

1. Address core educational needs of members
   - Assess member needs
   - Develop a "base-level knowledge of statistics" curriculum
   - Promote statistical engineering
   - Publish featured articles, special publications, and webinars
2. Build community and increase awareness by using diverse and effective communications
   - Webinars
   - Newsletters
   - Body of Knowledge
   - Web site
   - Blog
   - Social Media (LinkedIn)
   - Conference presentations (Fall Technical Conference, WCQI, etc.)
   - Short courses
   - Mailings
3. Foster leadership opportunities throughout our membership and recognize leaders
   - Advertise leadership opportunities/positions
   - Invitations to participate in upcoming activities
   - Student grants and scholarships
   - Awards (e.g. Youden, Nelson, Hunter, and Bisgaard)
   - Recruit, retain and advance members (e.g., Senior and Fellow status)
4. Establish and Leverage Alliances
   - ASQ Sections and other Divisions
   - Non-ASQ (e.g. ASA)
   - CQE Certification
   - Standards
   - Outreach (professional and social)

Updated October 19, 2013

Message From the Chair *Continued*

- Opportunities to learn and contribute by joining a Discussion Group or Interest Group at https://my.asq.org/communities/discuss/177/323
- Much, much more!

Looking ahead to 2021, we are doing some preliminary planning for these events (so mark your calendars!):

- **World Conference on Quality & Improvement WCQI 2021**—Anaheim, California from May 23–26. Visit wcqi.asq.org for more information. At the conference, please stop by the Statistics Division booth at the Expo to see us!
- **ASQ/ASA Fall Technical Conference (FTC) 2021**—Park City, Utah where we will offer short courses and conference sessions from October 13–15. When released, additional details will be at falltechnicalconference.org. Award winners for both 2020 and 2021 will be recognized.

Are you interested in playing a more active role in the Statistics Division? If so, contact chair@asqstatdiv.org. We are always looking for new volunteers who can help the division continue to grow. I look forward to meeting and working with many of you throughout the remainder of this year and into the next. As we continue pressing on, we will be stronger, wiser, and more equip in the end. Cheers!

Editor’s Corner

Harish Jose

Welcome to the second edition of ASQ Statistical Digest in 2020.

These are truly unprecedented times as we face a pandemic together. In fact, the term “unprecedented” hit its highest peak on Google Trends worldwide in April 2020. We want to thank all essential people working on the frontlines fighting this virus and working to keep us safe. Thank you!

We have another great edition this time including a timely article from Dr. Wheeler on Covid-19 data analysis. We have:

- Youden Address by James J. Filliben - The Role of DEX & EDA for Standards & the Role of Standards for DEX & EDA Part 2
- Statistical Process Control Column by Donald J. Wheeler - Covid-19 Data and Process Behavior Charts
- Hypothesis Testing Column by Jim Frost - Understanding Significance Levels
- Risk and Uncertainty Column by Stephen Luko—Interval Estimation
- Mini Paper by Melvin Alexander - Statistical Model Comparison Predicting Signs of Penetrating Abdominal and Pelvic Injuries using R
- Upcoming Conference Calendar

Youden Address: The Role of DEX & EDA for Standards & the Role of Standards for DEX & EDA Part 2

James J. Filliben., National Institute of Standards and Technology.

Introduction

Part 1 of this Youden story appeared in the February 2020 edition of the ASQ Statistics Digest (Vol. 39, No. 1, p. 5-19.)-https://my.asq.org/communities/files/177/5133. It focused on "The Youden Years" and provided historical insight into Jack Youden's career (pre-NBS and NBS), and showed how critical he was for the early growth of the statistical consulting group (SEL) at NBS/NIST. Youden's NBS career was a manifestation of how DEX (Design of Experiments) and EDA (Exploratory Data Analysis) could impact not only standards, but also science, engineering and industry. Youden set inspirationally high standards for the art and craft of statistical consulting, methodology development, and communication that carry on even to today.

Part 1 had 6 sections:

1. Youden: His Contributions, and JJF Personal Recollections

   This introduced Youden as the master consultant, experiment designer, data analyst, writer, and orator; as well as my early recollections of Youden as a then (1969) mostly-retired colleague.

2: The R. A. Fisher, NBS/SEL, & Jack Youden Connection: Historical Insights

   This discussed the common thread that R.A. Fisher's classic text "Statistical Methods for Research Workers" played in connection with the 1947 founding of SEL, the choice of SEL's first division chief (Churchill Eisenhart), and the choice of Jack Youden as an early SEL hire.

3: Youden Contributions at NBS

   This discussed Youden's experimental prowess, his love of applications and problem-solving, his prolific methodology contributions, and his masterful written and oral presentations.

4: Youden, NBS, & an Institutional Pivot Point: AD-X2

   This discussed Youden's expert response to the institutional challenge presented by the 1953 AD-X2 event, which was an existential threat to NBS as an independent scientific research laboratory. It also established

   SEL as a critically important component in NBS's mission of standards development and research.

5: Youden Chronology

   This outlined Youden's incredibly productive career from his early pre-NBS days as a research chemist, to his Fisher/Hotelling-led transformation into a statistical scientist, to his flourishing NBS days as experiment design expert, data analysis expert, master consultant and world-class communicator.

6: Youden: the Author & Communicator

This described Youden's productivity as an author: 114 publications (including 5 books) overall, across a variety of disciplinary platforms, and highlighted an incredibly productive NBS stretch 1959–1963 in which he had 41 publications (93% solo-authored!). Further it was noted that career-wise, Youden had 200+ lectures-universally characterized as clear, jargon-free and effective-which led to him being heavily in demand and acclaimed as a master oral communicator.

Part 2 (The "Post-Youden Years") here describes Youden's legacy and impact in the 5 decades since his death in the early 1970's. It covers specifics about the institutional evolution of the NBS Statistical Engineering Laboratory post-Youden, as well as methodological developments/ standardizations that (we contend) would hopefully find Youden's posthumous approval. Further, whereas Part 1 discussed Youden's life in the context of DEX & EDA for Standards, Part 2 discusses Youden's legacy in the context of Standards for DEX & EDA. In particular, Part 2 has 2 sections:

1. Post-Youden: NBS/SEL

   This summarizes the ongoing Youden effect/legacy on SEL over the past 5 decades. It enumerates SEL ASQ Youden Award Winners (4), ASA Youden Award winners (10), high profile SEL projects, SEL software, and SEL DEX workshops (24) (along with an associated standardized DEX teaching tool for 2-level design construction and confounding).

2. Post-Youden: Standardized DEX/EDA Tools: 4 Recommendations

   Inspired by Jack Youden (and John Tukey), this enumerates "standards" in DEX and EDA methodologies, and make 4 specific recommendations (one recommendation for each of 4 problem type) for such standards. We assert that just as standards in scientific research maximize accuracy; standards in DEX/ EDA construction maximize insight.

1: Post-Youden: NBS/SEL

It was seen in Part 1 that with the fortuitous R.A. Fisher-based trifecta of:

1. the fourth NBS Director Edward Condon,
2. the first Chief of the NBS Statistical Engineering Laboratory Churchill Eisenhart,
3. and the energetic and talented chemist-turned statistician Jack Youden

was critical to the establishment and growth of the Statistical Engineering Laboratory (SEL) within NBS. Here in Part 2, we provide a summary as to how the 50 post-Youden years have seen a continuation of the Youden spirit and a growth of NBS/SEL-even among changing standards demands, changing scientific problem arenas, and changing metrology tools in the hands of the NBS scientist. In contrast to the 5 members of SEL when Youden first joined in 1948, SEL currently has about 30 staff members-shared between the Gaithersburg MD campus (~3500) and the Boulder, CO campus (~600). The post-Youden SEL remains highly-respected and in 2022 will celebrate its 75th year as the statistical design and analysis consulting group here at NBS/NIST. We share here a few selected and abbreviated items about SEL in the post-Youden era-as a reaffirmation that the Youden legacy carries on in SEL and NIST in honor of the highest of professional standards that Youden has set for us all.

1.1: SEL Staff, Awards & Projects

SEL Division Chiefs: To date, SEL has had 13 chiefs. Many names are familiar and have had outstanding statistical careers both within NBS & SEL, and also outside of NBS & SEL:

Churchill Eisenhart (1947–1963); Joe Cameron(63–68); Joan Rosenblatt (68–78); Harry Ku (78–85); Mary Natrella (85–86) ; Bob Lundegard (86–94); Carroll Croarkin (acting) (94–95); Lynne Hare (95–96); Carroll Croarkin (96–97); Keith Eberhart (acting) (97–98); Keith Eberhart (98–99); Barbara Guttman (99–00); Nell Sedransk (2000–2005); Kamie Roberts (05–06); Antonio Possolo(06–13); Will Guthrie (13-present).

ASQ Youden Award Winners: Since 1973, the ASQ/ASA has hosted the Youden Address at the Fall Technical Conference. 4 members of SEL (or pre-SEL, or post-SEL) have been given the honor to present this address:

Churchill Eisenhart (1975); Brian Joiner (1978); Brian Joiner (1984); Lynne Hare (1993); Jim Filliben (2019).

ASA Youden Award Winners Since 1985, the American Statistical Association has annually presented the W. J. Youden Award in Interlab Testing for best paper in that area. Because interlab experimentation is a common problem arena at NBS, NBS/SEL has had many interlab articles over the years-10 of which have resulted in Youden Awards. SEL (or ex-SEL, or NBS/non-SEL) awardees include:

John Mandel (non-SEL, but still NBS) (1988); Cliff Spiegelman (post-NBS) (1991); John Mandel (non-SEL, but NBS) (1996); Andrew Rukhin #1 & Mark Vangel (1998); David Duewer, Margaret Kline, Katherine Sharpless, & Jeanice M. Brown Thomas (non-SEL, but still NBS) (2000); Jim Filliben (2003); Hari Iyer, Jack Wang, and Thomas Mathew (2005); Andrew Rukhin #2 with Bill Strawderman (2008); Blaza Toman (2009); Andrew Rukhin #3 (2018).

Visiting Staff

Churchill Eisenhart set the early tone for SEL research by assuring that many high-quality visitors would come through NBS/SEL. Post-Youden, SEL carried on this tradition of accomplished sabbatical visitors. Here is an abbreviated list:

1970s: Ray Sansing, John LeBrecque, Wes Nicholson, John Orban, Richard Jones, Jerry Sacks.

1980s: Karen Kafadar, John Rice, Ray Carroll, Jim Crichton, Dave Herbert, Nancy Flournoy, Sam Saunders, Ken Wallenius, Jim Williams, Stephen Sanulus, Leon Gleser, and Bobby Mee.

1990s: Wayne Nelson, Necip Doganasksoy, Mike Frey, Janos Galambos, Ker-Chau Lee, Sabri Cetikunt, Thad Tarpey, Moshe Pollack, Doug Simpson, Ali Cinar, Gene Hwang, Ghanashyam Joshi, Duane Boes, Jauarum Sethuraman, Yidaya Sivthanu, Nozer Singpurwalla. Bob Easterling.

Post-Youden SEL staff members who have contributed much to NBS and have then gone on to apply their experience and expertise at other institutions would include (for example): Cliff Spiegelman, Charlie Reeve, Grace Yang, Lisa Gill, Eric Lagergren, Keith Eberhart, Lynne Hare, Mark Levinson, Mark Vangel, and David Banks.

High-Visibility Projects: NBS/SEL/Gaithersburg has about 25 statisticians + our Boulder branch has about 5 statisticians. On a yearly basis, an SEL statistician will be involved in anywhere from as few as 6 to as many as 40 projects-some long term and many short term. A selective sample of 15 high-visibility projects since SEL's inception includes:

AD-X2, Selective Service Draft Lottery, Daylight Saving Time, Alaska Pipeline, Bullet-Proof Armor Testing, World Trade Center Collapse, DHS Radiation Detection, Video Analytics, NIST mAB Monoclonal Antibodies SRM 8671, Newton's Gravitational Constant G, Cell Biometrology, Net-Zero House Energy Consumption, and Cloud Computing Resource Allocation, and Gulf of Mexico Deepwater Horizon Oil Disaster.

1.2: Publications

In the spirit of Youden to produce relevant publications which address real-world metrology problems with rigorous, state-of-the-art methodologies, SEL staff have produced thousands of publications over the 5 decades-across a myriad of interdisciplinary journals. As for books, Youden had 5 [1951], [1960], [1962], [1967], [1974]. Post-Youden, there have been a number of noteworthy NBS/SEL books which (we contend) Youden himself would have been particularly "proud of"; here are 3:

NBS Handbook 91: Experimental Statistics) [Natrella, 1963]

Though not strictly speaking "post-Youden", it was published within a few years of Youden's formal retirement. This book by Mary Natrella (a long-standing colleague of Jack Youden) was very clearly written, very practical, and very highly regarded by the scientific and engineering community. It was the 2nd most published book ever at NBS and has been translated into 6 languages.

NBS Special Publication 300: Precision Measurement and Calibration-Statistical Concepts and Procedures [Ku, 1969] https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nbsspecialpublication300v1.pdf

This book by Harry Ku-also a long-standing colleague of Jack Youden-was/is an excellent collection of state-of-the-art statistical metrology methods. SP300 is unique in its collection and rigorous discussion of a wide range of fundamental metrology topics.

NIST/SEMATECH e-Handbook of Statistical Methods [Croarkin et al, 2003]

This 3000-page electronic document-authored by colleague Carrol Croarkin, 4 NIST SEL staff members, and 4 SEMATECH (the Semiconductor manufacturing consortium) staff members-is a 6-year joint effort of SEMATECH and NIST. It was first released in 2003. It was inspired and viewed as a modern update to Mary Natrella's NBS Handbook 91: Experimental Statistics. Both the Natrella book and the subsequent e-Handbook are characterized by:

- presenting real-world metrology problems,
- describing in detail state-of-the-art stat/EDA solutions,
- with many worked examples,
- using scientist-friendly terminology.

The 8 chapters are 1: Explore (EDA), 2. Measure (Measurement Process Characterization), 3. Characterize (Product Process Characterization, 4. Model (Regression), 5. Improve (Design of Experiment), 6. Monitor (SPC), 7. Compare (Hypothesis Testing), and 8. Reliability.

As a measure of its Youden-like utility, this e-Handbook:

1. has been viewed ~ 650,00 to 800,000 per year for the last 15+ years, and
2. has served as Google's top- (or 2nd top-) hit for many common statistical terms (e.g., Youden plot, probability plot, hazard plots, contour plots, ppcc plots, box-cox normality plots, bihistograms, block plots, consensus-mean plots, etc.)

To access the e-Handbook: http://www.itl.nist.gov/div898/handbook/

To access its useful statistical graphics gallery: https://www.itl.nist.gov/div898/handbook/graphgal.htm

This gallery is ordered by common problem category, with a variety of graphical methodologies assembled under each category. Figure 5 shows a page from the Univariate problem category.

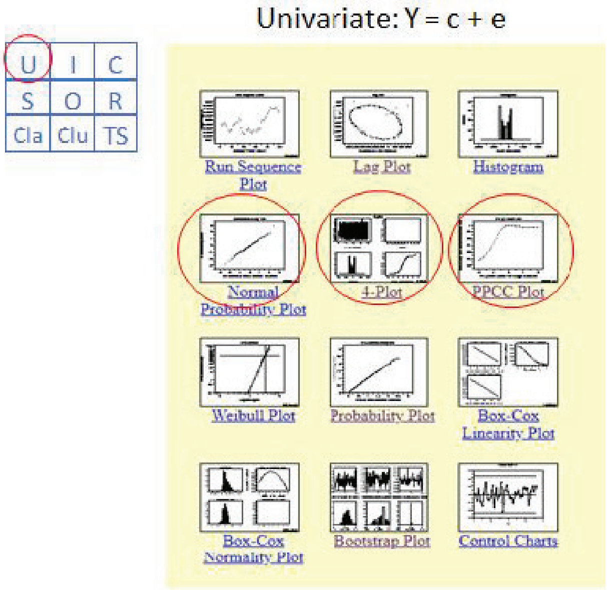


Figure 5: e-Handbook Graphics Gallery, page 1 of Univariate

1.3: Statistical Software

Youden was the consummate problem-solver and tool developer. In that light, note that SEL has had 2 major contributions to the statistical software tools arena, both of which have played key roles in the field of early state-of-the-art statistical computing:

1. **Omnitab:** This system was first released in 1968 and was headed by SEL colleague Dave Hogben with assistance by SEL's Sally Peavy, Ruth Varner, and Shirley Bremer. When Omnitab was released, there were only 5 stat software systems in the world: including BMDP (UCLA) and Roald Buhler's P-Stat (Princeton). In the early 1970's, Omnitab was carried from NBS by my colleague Brian Joiner and was modified by Brian and the (Tom & Joan) Ryans to serve as a Penn State in-class computational teaching tool. Eventually it morphed into Minitab, which of course is still in heavy use today.
2. **Dataplot:** This public-domain system was developed and first released in 1978 by Jim Filliben. It was designed for interactive graphics and non-linear fitting, and it was the most popular public-domain software system for such in the early 1980's, with many non-NBS installations across industry and government. It was the first statistical software system presented at the prestigious ASTM ACM SIGGRAPH Computer Graphics '81 Conference [Filliben, 1981]. Along with SEL's Alan Heckert, Dataplot still carries on today with decades' worth of standardized EDA graphical tools-in the spirit of Youden (and Tukey). One such tool, for example, uses the Youden Plot and applies it to the analysis of 2-level orthogonal designs. Other tools include 4-plot univariate analyses, PPCC (Probability Plot Correlation Coefficient) plots, box-cox normality plots, consensus-value plots, block plots, standardized 10-step EDA analysis of orthogonal 2-level sensitivity analysis designs, etc. All graphics in this Youden manuscript are Dataplot-generated.

For graphics commands: https://www.itl.nist.gov/div898/software/dataplot/refman1/ch2/homepage.htm

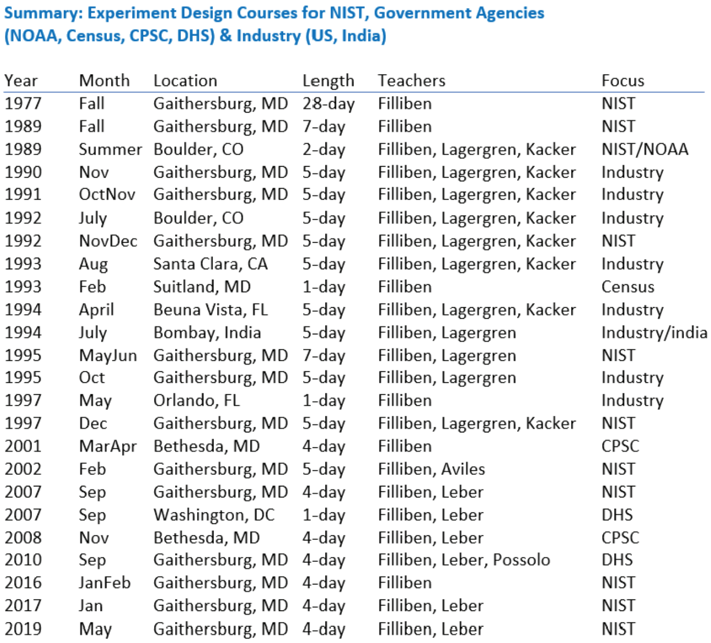
For analysis commands: https://www.itl.nist.gov/div898/software/dataplot/refman1/ch3/homepage.htm

For homepage, graphics gallery, & free downloading: http://www.itl.nist.gov/div898/software/dataplot/

1.4: DEX Workshops

Jack Youden gave a remarkable 200+ lectures and workshops over his career. Many of those talks focused on principles and techniques of experiment design. In a similar fashion (and with additional inspiration from my Princeton professor Stu Hunter with his own equally remarkable career in creating/teaching/applying experiment design), it became apparent that efficient and rigorous experimental plans were critical to research success for the NBS/NIST scientist/engineer. In this context, DEX training has been an ongoing priority for NBS/NIST/SEL over the last 5 decades. Figure 6 presents a history (from 1977 to 2019) of SEL DEX training given both within NBS and outside of NBS in the post-Youden SEL-era:

Jack Youden
(Photo: Public Domain)



Summary: Experiment Design Courses for NIST, Government Agencies (NOAA, Census, CPSC, DHS) & Industry (US, India)

Year	Month	Location	Length	Teachers	Focus
1977	Fall	Gaithersburg, MD	28-day	Filliben	NIST
1989	Fall	Gaithersburg, MD	7-day	Filliben	NIST
1989	Summer	Boulder, CO	2-day	Filliben, Lagergren, Kacker	NIST/NOAA
1990	Nov	Gaithersburg, MD	5-day	Filliben, Lagergren, Kacker	Industry
1991	OctNov	Gaithersburg, MD	5-day	Filliben, Lagergren, Kacker	Industry
1992	July	Boulder, CO	5-day	Filliben, Lagergren, Kacker	Industry
1992	NovDec	Gaithersburg, MD	5-day	Filliben, Lagergren, Kacker	NIST
1993	Aug	Santa Clara, CA	5-day	Filliben, Lagergren, Kacker	Industry
1993	Feb	Suitland, MD	1-day	Filliben	Census
1994	April	Beuna Vista, FL	5-day	Filliben, Lagergren, Kacker	Industry
1994	July	Bombay, India	5-day	Filliben, Lagergren	Industry/india
1995	MayJun	Gaithersburg, MD	7-day	Filliben, Lagergren	NIST
1995	Oct	Gaithersburg, MD	5-day	Filliben, Lagergren	Industry
1997	May	Orlando, FL	1-day	Filliben	Industry
1997	Dec	Gaithersburg, MD	5-day	Filliben, Lagergren, Kacker	NIST
2001	MarApr	Bethesda, MD	4-day	Filliben	CPSC
2002	Feb	Gaithersburg, MD	5-day	Filliben, Aviles	NIST
2007	Sep	Gaithersburg, MD	4-day	Filliben, Leber	NIST
2007	Sep	Washington, DC	1-day	Filliben, Leber	DHS
2008	Nov	Bethesda, MD	4-day	Filliben, Leber	CPSC
2010	Sep	Gaithersburg, MD	4-day	Filliben, Leber, Possolo	DHS
2016	JanFeb	Gaithersburg, MD	4-day	Filliben	NIST
2017	Jan	Gaithersburg, MD	4-day	Filliben, Leber	NIST
2019	May	Gaithersburg, MD	4-day	Filliben, Leber	NIST

Figure 6: NBS/SEL DEX Workshops

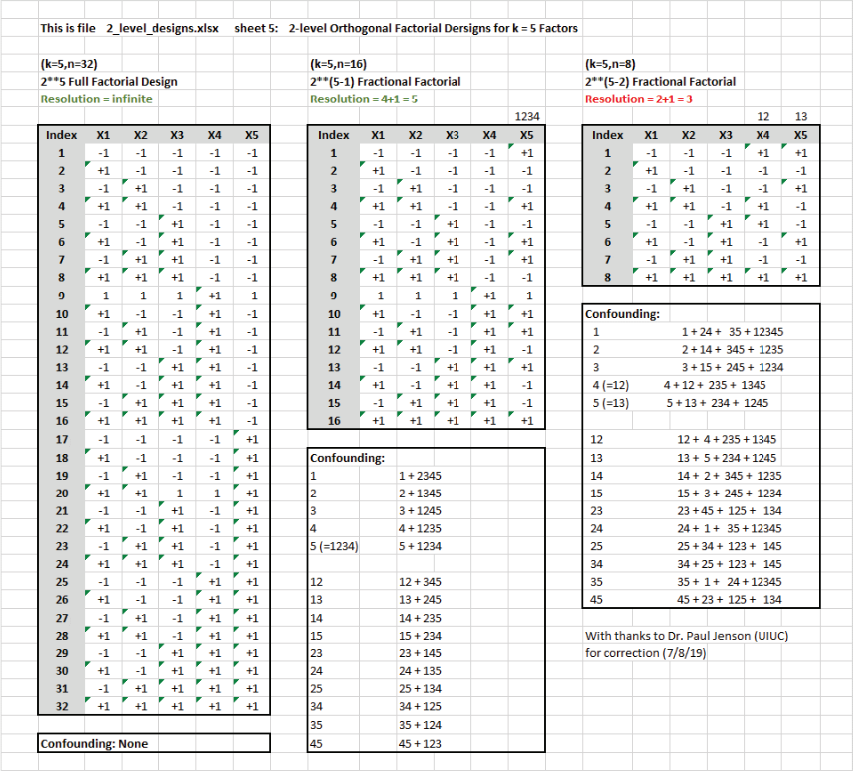
1.5: DEX: Standardized Online Access to 2-Level Designs

With Youden's example-by-doing, and with methodological roots going all the way back to Fisher, DEX/DOE quickly became an essential component in what the SEL statistical consultant offers the NBS/NIST researcher. In particular, over the decades, it has become increasingly obvious as to the central role that orthogonal 2-level factorial designs play in efficiently addressing the sensitivity/screening analysis problems that NBS researchers encounter daily. For a given (k factor, n run) experiment, two immediate issues arise for the researcher:

1) how does one construct the design, and

2) how does one determine the confounding structure for the fractionals.

In this regard, and in the additional regard of having an auxiliary in-class tool to assist in the aforementioned DEX Workshops, this led to the design and construction of a single, multi-tabbed, Excel file which serves as a handy one-source repository for all of the 2-level orthogonal full and fractional factorial designs (along with confounding structure) in the most frequently-encountered factor domain (k = 2, 3, 4, . . . , 11). For many of the designs, credit is due ultimately to the pioneering work of Box, Hunter, & Hunter (see p. 410 of their original classic text "Statistics for Experimenters" (1978), and p. 272 of their edition 2 (2005)). The designs (and confounding) for a given number of factors k are accessible by selecting tab k of this worksheet file). For example, Figure 6-accessed by selecting tab 5 of this excel file-shows all of the full and fractional designs for k = 5 factors. This design file has been transferred to ASQ and may be accessed at https://my.asq.org/communities/files/177/5731



This is file 2_level_designs.xlsx sheet 5: 2-level Orthogonal Factorial Dersigns for k = 5 Factors

(k=5,n=32)
2**5 Full Factorial Design
Resolution = infinite

Index	X1	X2	X3	X4	X5
1	-1	-1	-1	-1	-1
2	+1	-1	-1	-1	-1
3	-1	+1	-1	-1	-1
4	+1	+1	-1	-1	-1
5	-1	-1	+1	-1	-1
6	+1	-1	+1	-1	-1
7	-1	+1	+1	-1	-1
8	+1	+1	+1	-1	-1
9	1	1	1	+1	1
10	+1	-1	-1	+1	-1
11	-1	+1	-1	+1	-1
12	+1	+1	-1	+1	-1
13	-1	-1	+1	+1	-1
14	+1	-1	+1	+1	-1
15	-1	+1	+1	+1	-1
16	+1	+1	+1	+1	-1
17	-1	-1	-1	-1	+1
18	+1	-1	-1	-1	+1
19	-1	+1	-1	-1	+1
20	+1	+1	1	1	+1
21	-1	-1	+1	-1	+1
22	+1	-1	+1	-1	+1
23	-1	+1	+1	-1	+1
24	+1	+1	+1	-1	+1
25	-1	-1	-1	+1	+1
26	+1	-1	-1	+1	+1
27	-1	+1	-1	+1	+1
28	+1	+1	-1	+1	+1
29	-1	-1	+1	+1	+1
30	+1	-1	+1	+1	+1
31	-1	+1	+1	+1	+1
32	+1	+1	+1	+1	+1

Confounding: None

(k=5,n=16)
2**(5-1) Fractional Factorial
Resolution = 4+1 = 5

					1234
Index	X1	X2	X3	X4	X5
1	-1	-1	-1	-1	+1
2	+1	-1	-1	-1	-1
3	-1	+1	-1	-1	-1
4	+1	+1	-1	-1	+1
5	-1	-1	+1	-1	-1
6	+1	-1	+1	-1	+1
7	-1	+1	+1	-1	+1
8	+1	+1	+1	-1	-1
9	1	1	1	+1	1
10	+1	-1	-1	+1	+1
11	-1	+1	-1	+1	+1
12	+1	+1	-1	+1	-1
13	-1	-1	+1	+1	+1
14	+1	-1	+1	+1	-1
15	-1	+1	+1	+1	-1
16	+1	+1	+1	+1	+1

Confounding:	
1	1 + 2345
2	2 + 1345
3	3 + 1245
4	4 + 1235
5 (=1234)	5 + 1234
12	12 + 345
13	13 + 245
14	14 + 235
15	15 + 234
23	23 + 145
24	24 + 135
25	25 + 134
34	34 + 125
35	35 + 124
45	45 + 123

(k=5,n=8)
2**(5-2) Fractional Factorial
Resolution = 2+1 = 3

				12	13
Index	X1	X2	X3	X4	X5
1	-1	-1	-1	+1	+1
2	+1	-1	-1	-1	-1
3	-1	+1	-1	-1	+1
4	+1	+1	-1	+1	-1
5	-1	-1	+1	+1	-1
6	+1	-1	+1	-1	+1
7	-1	+1	+1	-1	-1
8	+1	+1	+1	+1	+1

Confounding:	
1	1 + 24 + 35 + 12345
2	2 + 14 + 345 + 1235
3	3 + 15 + 245 + 1234
4 (=12)	4 + 12 + 235 + 1345
5 (=13)	5 + 13 + 234 + 1245
12	12 + 4 + 235 + 1345
13	13 + 5 + 234 + 1245
14	14 + 2 + 345 + 1235
15	15 + 3 + 245 + 1234
23	23 + 45 + 125 + 134
24	24 + 1 + 35 + 12345
25	25 + 34 + 123 + 145
34	34 + 25 + 123 + 145
35	35 + 1 + 24 + 12345
45	45 + 23 + 125 + 134

With thanks to Dr. Paul Jenson (UIUC)
for correction (7/8/19)

Figure 7: Excel File, Tab 5 (for 2^5 and 2^{5-p} Designs & Confounding)

Upon reflection, it is indeed sobering (to this writer) to consider that all of these post-Youden NBS/SEL contributions owe their existence to a single book (Fisher's 1925 Statistical Methods for Research Workers), and to the 3 people (Condon, Eisenhart, and Youden) who had made it a priority to read that book. In the absence of those 3 interconnected people and events, it is entirely plausible that subsequent events-including SEL's establishment/existence/contributions and NBS's unblemished tenure as the nation's metrology laboratory of last resort-would not have come to pass.

Summary-Post-Youden: NBS/SEL

Upon reflection, it is indeed sobering (to this writer) to consider that all of these post-Youden NBS/SEL contributions owe their existence to a single book (Fisher's 1925 "Statistical Methods for Research Workers"), and to the 3 people (Condon, Eisenhart, and Youden) who had made it a point to read that book. In the absence of these 3 interconnected people and events, it is entirely plausible that subsequent events-including SEL's establishment/existence/contributions and NBS's unblemished tenure as the nation's metrology laboratory of last resort-would not have come to pass. On a more positive note, the "bottom line" for this entire Post-Youden NBS/SEL section 1 (awards, pubs, software, workshops, etc.) is of course obvious: the spirit of Youden is still alive, well, and pervasive in SEL at NBS/NIST-50+ years later.

2. Post-Youden: Standardized DEX/EDA Tools: 4 Recommendations

The 1969 Wilks Awards to Jack Youden states:

> ". . . for his extensive contributions to the art and practice of experimentation in the sciences and engineering through conception and lucid exposition of novel, rather elementary techniques of statistical analysis and crafty application of standard methods; and through his . . . indefatigable energy and phenomenal effectiveness as a speaker . . ."

We here highlight the phrase "novel, rather elementary techniques of statistical analysis" (which focuses on Youden as a Stat/EDA tool developer) and link it with the second half of this Youden Address' title: "The Role of DEX & EDA for Standards, & the Role of Standards for DEX & EDA". It is my belief that for certain categories of problems, standard EDA methodologies should exist and should be routinely applied whenever such problem-types arise. Just as calibration standards in science are critical for accuracy; methodology standards in data analysis are critical for insight. In the spirit, for example, of the Youden Plot being a "standard" powerful tool for interlab problems, we would like to recommend 4 such similar (in spirit) "standard" tools (EDA graphical analysis methodologies) that have been developed over the years. These are routinely used to provide powerful insight for the following 4 problem types at NBS/NIST (with obvious applications beyond to science, engineering, and industry):

1. Univariate: 4-plot
2. Interlab: consensus-value plot
3. Comparative: block plot
4. Sensitivity Analysis: DEX 10-step analysis

We use this Youden Address forum to share them because of their similarity in spirit to both the Jack Youden (and the John Tukey) approach for extracting information and insight from data. Because of space considerations, we present only a limited discussion here; for further reading, see:

1. Chapter 1 of the NIST/SEMATECH e-Handbook of Statistics: http://www.itl.nist.gov/div898/handbook/,

2. the e-Handbook's graphics gallery: https://www.itl.nist.gov/div898/handbook/eda/section3/eda33.htm,
3. as well as Dataplot's graphics gallery: https://www.itl.nist.gov/div898/software/dataplot/gra_gall/homepage.htm.

We (biasedly) believe/hope these 4 tools would receive "Youden's blessing" in both spirit and substance.

2.1 Standardized Tool #1: 4-Plot for Univariate Problem

The univariate problem type is the simplest one-we have a column of numbers. What can we say about them? The underlying model for this problem is:

$Y = c + e;$

where c is a typical value and e is an error term with (location 0, some scale, and some distribution). The assumption is often made that the observations are independent of one another.

The generic starting point for our standardized analysis (the "4-plot") is driven by the question:

Q. Is this process in statistical control?

That is, are the numbers emanating from the "process" behaving like:

1. random numbers
2. from a fixed distribution,
3. with a fixed location, and
4. with a fixed scale.

With an implicit fifth item to this list being

5. with no outliers

If the above 5 items hold, then we made state that the process is "in statistical control" (or more rigorously, that there is no evidence from the data that the process is "out of statistical control"). If the process is deemed "in control", then we may advantageously (but cautiously) make probability statements not only about where the process has been, but also about where the process will be in the future. We may thus go from mere local data summarization to global data inference.

In response to these 4 questions, we recommend the "4-plot" consisting of the following plots are for "standard" use:

1. run sequence plot: Yi vs dummy index i
2. lab plot: Yi vs Yi-1
3. histogram
4. normal probability plot (ordered Yi vs order stat medians from a N(0,1) distribution)

In figure 8 we apply the 4-plot analysis to an ideal: simulated normal N(100,10) data. Interpretationally from the figure, we note that the run sequence plot has no drift and has a fixed variation band, the lag plot has no structure, the histogram is bell-shaped, and the normal probability plot is linear. Further, the linearity of the normal probability plot is quantitatively affirmed by the normal probability plot correlation coefficient (see Filliben [1975]). Note also that no outliers are present. We admit that this 4-plot analysis does not formally test as to whether the distribution is fixed per se; rather, it does alternatively provide feedback as to what distribution the data has-an informative starting point. In summary, we would conclude from figure 8 that this process is "in control" (as it should be since these numbers were random by construction). Thus in the spirit of elementary (but insightful) graphics espoused by Youden, we posit that the 4-plot method would fall into that category. In practice, the 4-plot is routinely used as a "standard" first pass (necessary, but not always sufficient) tool for any univariate data.

The second 4-plot example (figure 9) is from 200 observations drawn from an NBS/NIST Center for Building Technology beam deflection process. What can be said about this data and the process behind the data? Interpretation-wise, we see that the run sequence plot shows no drift, and has a fixed variation band (with a hint of a single outlier around observation 160), the lag plot is rather strikingly elliptical and has 4 off-elliptic points (generated from 2 outliers), the histogram is bimodal, and the normal probability plot fails. From this 4-plot, we would conclude -especially from the lag plot Lissajous-like pattern-that the underlying process is cyclic, with not just 1-but 2-outliers.

We find the 4-plot to be a useful check of any univariate data, and any residuals (regression, ANOVA, etc.) from any model. For more details, see

1. e-Handbook: https://www.itl.nist.gov/div898/handbook/eda/section3/4plot.htm
2. Dataplot Graphics Gallery: https://www.itl.nist.gov/div898/software/dataplot/gra_gall/4-plot.htm
3. Dataplot 4-Plot Command: https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/4-plot.htm

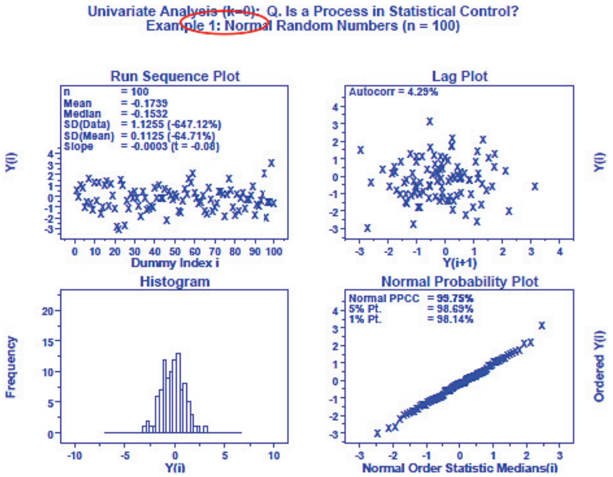


Figure 8: 4-plot for Normal Data (n = 100)

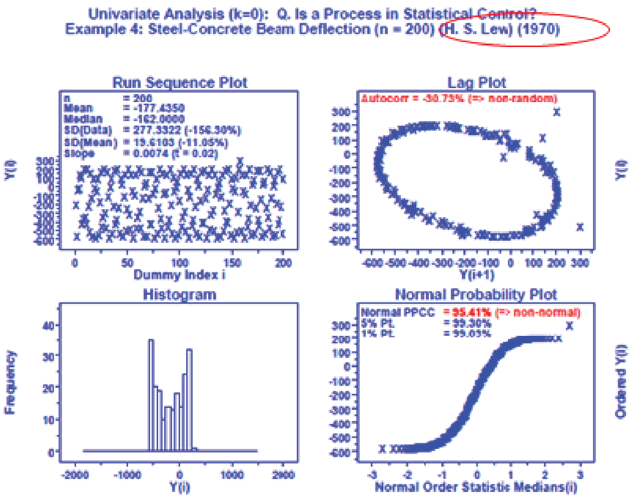


Figure 9: 4-plot for Beam Deflection Data (n = 200)

Note: A higher resolution image is available at https://my.asq.org/communities/files/177/5730

2.2 Standardized Tool #2: Consensus Value Plot for Interlab Problem

For the second tool, we look at another important NBS/NIST standardization problem: the interlab consensus problem: given data from many "expert" labs on a common reference material, what consensus value should be used as the "certified value" that goes on the institutionally official NIST SRM (Standard Reference Material) Certificate-and with what uncertainty? The model is:

$Y = c + f(\text{lab}) + e$

where ideally the f(..) function is null, but in reality may be non-null due to (small or large) interlab procedural or environmental differences. In spite of the unknown f(..), what should the consensus estimate be for c, and with what uncertainty?

The solution we use and share is the Consensus Value Plot (see figure 10). It consists of 2 subplots. The left subplot is a plot of the raw data (vertically) versus the lab ID (horizontally). The right subplot is a plot of consensus value estimates (vertically) versus estimation method (horizontally). To estimate a consensus value from data, we find that rather than attempting to determine the optimal statistical consensus value estimator by ascertaining as to which assumptions from which estimators hold for this dataset, we rather try a different approach by computing a battery of 13 different consensus values from 13 different commonly-used statistical consensus estimators (mean of means, Mandel-Paule, Vangel-Rukhin, DerSimonian-Laird, median of means, midmean, trimmed mean, etc.) and display such estimators (and uncertainties) on the right subplot. In many cases, most-if not all-of the estimators are relatively near-equivalent in value and uncertainty. In practice, this is a useful starting-point robustness conclusion unto itself. At a very minimum, the multi-estimate plot and right-margin tabulation gives the researcher practical worst-case bounds on the consensus value and its uncertainty.

If the 13 estimates happen to be near-equivalent, then parsimony dictates that we use the simplest estimator (= mean of means). If they are not equivalent, we find ourselves recommending the simplest robust conservative estimator, namely, the median of means. In the majority of consensus-value data cases we have encountered, this conservative value has been more than "fit" for the standards scientist's "purpose at hand". The median of means is an especially good choice in negating the effect of some "outlying" labs. In any event, in the spirit of Youden's "novel, rather elementary techniques of statistical analysis", this Consensus Value Plot displays both the raw input data as well as the output collection of possible answers, and so serves as a good first (and sometimes last) pass for the consensus value problem.

As it turns out, the data we chose to use for figure 10 was from Youden's classic paper thickness problem [Youden, 1962] involving 24 high school students, and their measurement of the thickness of a paper page. Each student in effect became a "lab". The usual 4 consensus value questions arose:

1 Are the 24 students/ "labs" equivalent?
2. Any outlying students/ "labs"?
3. What is a consensus value?
4. What uncertainty?

It is clear from figure 10 that student 21 has a possible outlier (and Grubbs test could be used to confirm this), but nevertheless the simple median of means (here=0.07761 +- .00212) would again be a good outlier-independent robust choice for the consensus value and (k=2) uncertainty. Visually from the right subplot, this is similar in value to the mean of means and other classic estimators (with their more stringent assumptions). In any event, for this problem type, the Consensus Value Plot is an excellent starting point which gives not only outlier & robustness insight immediately, but also provides a superset of probable final consensus value estimates. This plot is most commonly used at NBS/NIST when the robustness factor is laboratory; but may of course be used in a broader setting for any factor (batch, vial, operator, day, etc.) which ideally should have no effect, but in reality may have a non-negligible contaminating effect. This ASQ Stat Digest serves as the first reference for the Consensus Value plot.

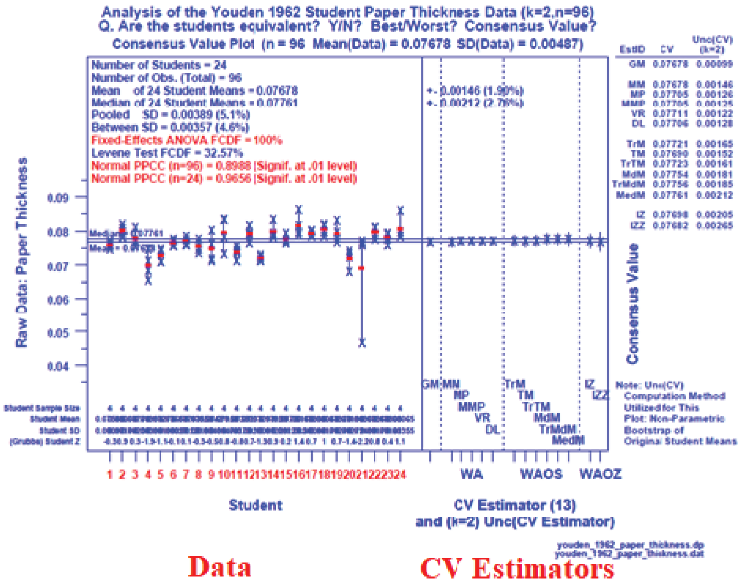


Figure 10: Consensus Value Plot for Youden Paper Thickness Data (n = 96)
Note: A higher resolution image is available at https://my.asq.org/communities/files/177/5732

2.3 Standardized Tool #3: Block Plot for Comparative Problem

Yet another classic NBS/NIST problem is where the response is some function of several factors, but the focus of the study is whether a particular (scientist-chosen) single factor (=the "primary" factor) is significant or not, and whether the significance conclusion robustly holds over all of the settings of all of the other factors. If the primary factor effect is consistent over all robustness factor settings, then our primary factor conclusion is robust; if not consistent, then an interaction exists in which case it must be identified and characterized. The model is:

$Y = f(X_p, X_{r1}, X_{r2}, X_{r3}, \ldots, X_{rk}) + e$

where the function f is unknown (and will remain unknown), the factor Xp is the primary factor for which a definitive effect statement is to be made, and where the Xr1, Xr2, etc. are the robustness factors which serve to embroaden the scope of our conclusions about Xp. At NBS/NIST, common primary factors are method, device, algorithm, operator, fabrication, etc. In fact, any factor may be chosen as a primary factor-it is up to the scientist do decide whether a particular factor is (or is not) of special scientific interest to him/her for this study, and then determine by subsequent experimentation as to whether that factor is in fact significant, and whether that significance conclusion is robust over all other factors.

The most common graphical method in practice for determining the significance of a single factor is the scatter plot. Such a plot may yield significance (via t tests or ANOVA), but unfortunately yields very little additional insight as to the robustness of that conclusion, and virtually no information about the existence and nature of possible interactions between the primary factor and some robustness factor.

As an example, let us draw on the classic 2^5 chemical reactor example in Box, Hunter, & Hunter [2005], pages 260–261]. This is a (k=5, n=32) factor full factorial experiment with response=% reacted, and the 5 factors are feed rate, catalyst, agitation rate, temperature, and concentration. For sake of demonstration, let us assume that the researcher is particularly interested in the effect of X2=catalyst, and so the title of this comparative study might be something like "Assessment of the effect of catalyst on chemical reactor efficiency" as opposed to the more usual multifactor sensitivity analysis title: "Determination of the most important factors affecting chemical reactor efficiency".

Figure 11 is a scatter plot of efficiency vs each of the k=5 factors. Focusing on the second factor (X2: catalyst) we have % reacted on the vertical axis vs the primary factor X2 placed as usual as the second item on the horizontal axis. From the plot, we conclude that catalyst is indeed significant (with catalyst 2 yielding % reacted values which are on average about 25 units higher than catalyst 1 (further, a t test or ANOVA would affirm that this primary factor is in fact statistically significant). That is the good news; the bad news is that the scatter plot does not have the ability to convey robustness information and interaction information.

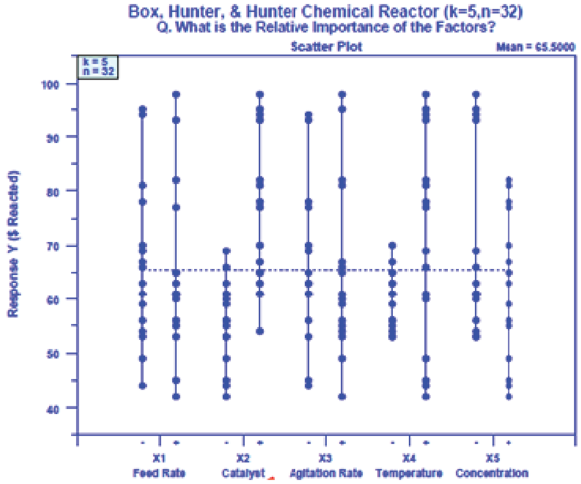


Figure 11: Scatter Plots for Box, Hunter, & Hunter Chemical Reaction Data (n = 32)
Note: Additional information is available at https://my.asq.org/communities/files/177/5729

An alternative and preferred analysis to the scatter plot is the block plot (figure 12). On the vertical axis is the mean % reacted; on the horizontal axis are all 16 combinations of settings for the 4 robustness factors. Where does primary factor information get conveyed? It is done via the plot character (1,2 = catalyst 1,2). The block plot thus takes the global question Q. Does catalyst have an effect? and decomposes it into 16 local questions: Q. Does catalyst have an effect for robustness condition 1? For robustness condition 2? . . . For robustness condition 16? Does catalyst have an effect for most conditions? For all conditions? If yes, then that yields a much stronger, much more compelling, and much more robust conclusion than a simple scatter plot. In fact, from figure 12, we see that for this example, catalyst 1 is indeed smaller than catalyst 2 for each and every one of the 16 conditions, and so by simple binomial (instead of t and ANOVA) considerations, we may confidently conclude that catalyst is not just statistically significant but is robustly statistically significant.

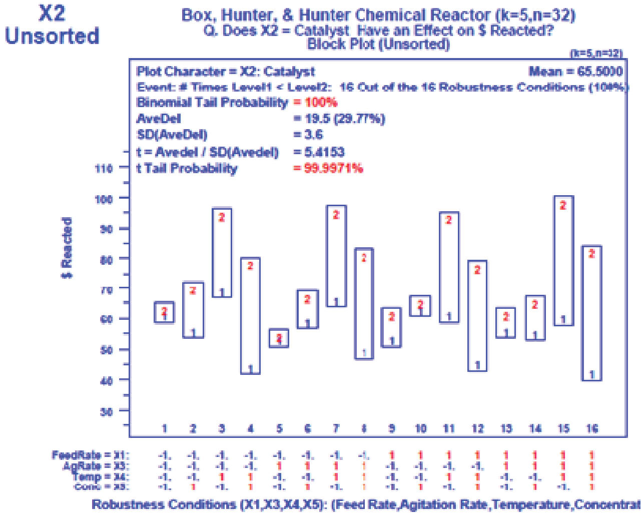


Figure 12: Block Plot for Box, Hunter, & Hunter Chemical Reaction Data (n = 32)
Note: Additional information is available at https://my.asq.org/communities/files/177/5728

Further (and even better) figure 13 shows a sorted normalized block plot. To construct this plot, the first step is to normalize each block in figure 12 by computing each of the 16 block means and then subtracting them out within each block. We thus have formed residuals-localized for each block. These residuals have "amplified" primary factor information. The resulting normalized plot would force us to focus on what is happening within a block (the primary factor effect) and not be distracted by what is happening between blocks (robustness factor effects). After such normalization, the next step is to note the 16 catalyst effects (here = the 16 block heights), and then simply sort the 16 blocks from smallest to largest and carry along the corresponding 16 horizontal axis robustness conditions. The net result is figure 13.

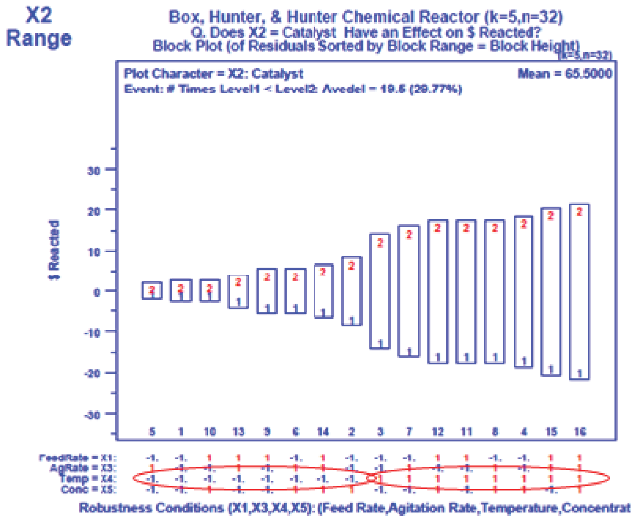


Figure 13: Sorted Normalized Block Plot for Box, Hunter, & Hunter Chemical Reaction Data (n = 32)

From figure 13: we see 3 items of interest:

1. we see that the catalyst effect is not constant over all 16 conditions; that is, there is an interaction (an existent interaction is one of the most important conclusions that can flow out of a comparative analysis);
2. we see clearly those robustness factor conditions which yield smaller catalyst effects, and those which yield larger catalyst effects. (This specific condition dependency is also useful to know).
3. by inspection of the conditions, we see the dominant robustness factor that drives the catalyst effect (the block size). For this example, we see (highlighted in red in figure 13) that the X2 catalyst effect is consistently small when factor X4 (temperature) is at its lower setting, and the X2 catalyst effect is consistently large when factor X4 is at its higher setting. This is a clear visual identification and confirmation that there is an X2*X4 (catalyst*temperature) interaction.

In the NBS/NIST environment, when a single factor has been chosen as the focus of the scientific study, then the block plot has been routinely used beyond the scatter plot to provide robustness and interaction information. Even stronger, it is our considered opinion and experience that the sorted normalized block plot is among the best of EDA techniques for discovering and understanding interactions, if existent, in comparative experiments. For more details, see

1. Filliben, J.J., Cetinkunt, S., Yu, W.Y., and Donmez, A. (1993)
2. e-Handbook: https://www.itl.nist.gov/div898/handbook/eda/section3/blockplo.htm
3. Dataplot : https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/blocplot.htm

2.4 Standardized Tool #4: DEX 10-Step for Sensitivity Problem

In the NBS/NIST environment, especially in the scientific research component, the most common scientific problem type by far is that of sensitivity/screening, whereby the scientist has a response of interest that is affected by many factors, and is interested in knowing what factors (and interactions) are important, and what are unimportant. The most popular and efficient NBS/NIST designs for this problem are the 2-level orthogonal full and fractional factorial designs. The primary deliverable for the sensitivity/

screening problem type is a ranked list of factors (and interactions)-with the sorting done based on the magnitude of the effect. Given that, a "free" secondary deliverable is a list of the best settings of the k factors, and a "free" tertiary deliverable is a formal empirical model.

The model for this sensitivity/screening problem type is:

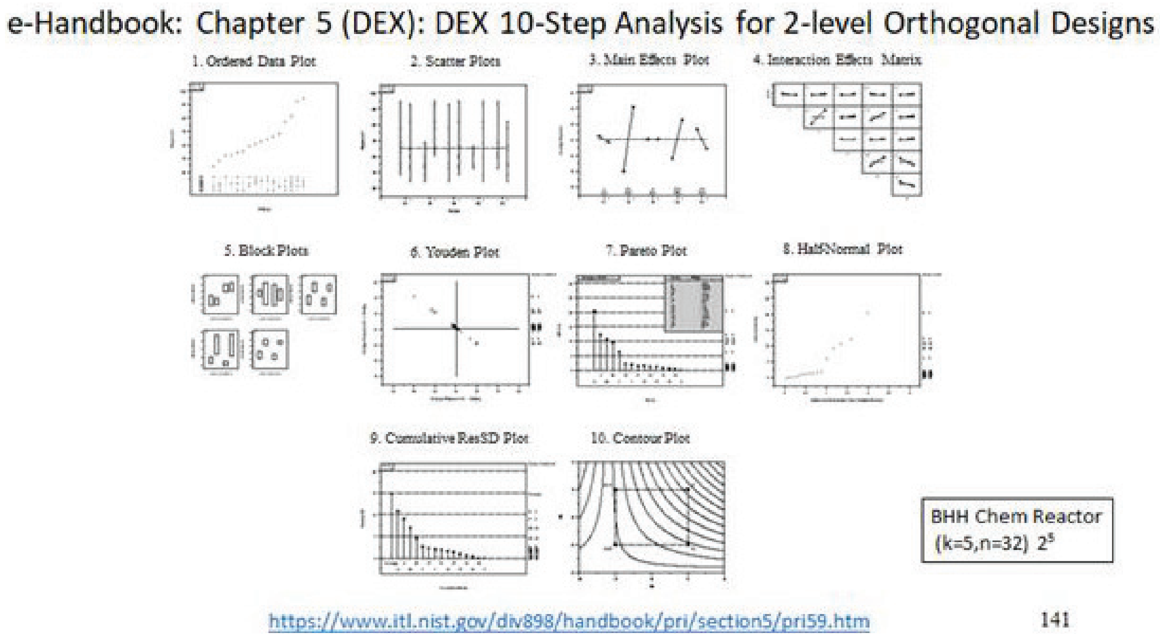
$Y = f(X_1, X_2, X_3, \ldots, X_k) + e$

where the function f is unknown, and X_1 to X_k are factors of equal scientific interest. Many important NBS/NIST projects at NIST have utilized 2^{k-p} designs, including the World Trade Center Collapse 2^{13-9} (k = 13 factors, n = 16 + 1 runs).

For this problem type and utilized 2-level designs, we have constructed over the years the following "standardized" 10-step procedure which has given us great insight into "understanding the system". In the spirit of Youden's "novel, rather elementary techniques of statistical analysis", we recommend the following "standardized" 10-step analysis procedure for the analysis or sensitivity/screening data via 2^k and 2^{k-p} orthogonal fractional factorial designs (along with the expected deliverable from each step):

1. ordered data plot: to determine best settings and most important factor
   => best settings & important factors
2. dex scatter plot: to determine most important factors (see, e.g., figure 11)
   => important factors
3. main effects plot: to determine most important factors
   => important facctors & best settings
4. interaction effects plot: to determine important 2-term interactions + derive confounding structure
   => important factors and interactions
5. dex block plots: to determine important factors and interactions
   => important factors and interactions
6. dex youden plot: to determine important factors and interactions
   => important factors
7. dex effects plot: to produce ranked list of important factors and interactions
   => ranked list of important factors and interactions
8. half-normal probability plot: to determine most important factors and interactions
   => important factors and interactions
9. cumulative residual SD plot: to assess goodness of fit of sequential models
   => good empirical model
10. contour plot: to assess an interaction & determine direction for future experiments
    => best factorial direction for next experiment

Space-considerations prevent a further detailed discussion of these 10 techniques; rather, we note that this parsimonious, standardized, 10-step graphical methodology exists for this problem type, and we present the following summary display (figure 14) which applies the 10-step analysis to the same Box, Hunter, & Hunter [2005, pages 260–261] classic 2^5 chemical reactor example as first mentioned in section 2.3, but treating all 5 factors as "equal" in terms of scientific interest, and so it becomes (as BHH intended) a traditional sensitivity/screening problem type.



e-Handbook: Chapter 5 (DEX): DEX 10-Step Analysis for 2-level Orthogonal Designs

BHH Chem Reactor
(k=5,n=32) 2^5

https://www.itl.nist.gov/div898/handbook/pri/section5/pri59.htm

141

Figure 14: DEX 10-Step Analysis for Box, Hunter, Hunter Chemical Reaction Data (n = 32)

For a detailed explanation of the construction and interpretation of the 10-step analysis, see https://www.itl.nist.gov/div898/handbook/pri/section5/pri59.htm

For high-resolution images for each of the 10 plots in figure 14, see https://my.asq.org/communities/files/177/5729

For the 10-step analysis applied to a 16-run sensitivity analysis of 8 factors affecting the error in a NIST HRTEM (High Resolution Tomographic Electron Microscope) along with the start-to-finish problem statement, design, confounding structure, data, script, conclusions, and the resulting published manuscript (in the physics journal Measurement Science & Technology [Scott, 2007]) see https://my.asq.org/communities/files/177/5733

Summary-Post-Youden: Standardized DEX/EDA Tools: 4 Recommendations

In summary, section 2 (Post-Youden: Standardized DEX/EDA Tools: 4 Recommendations) presents 4 standardized EDA methodologies that have evolved at NBS/NIST in the post-Youden era. The 4 tools (4-plot, consensus-value plot, block plot, and DEX 10-step analysis) have been selected to illustrate the ongoing spirit of Youden at NBS/NIST in pursuing "novel, rather elementary techniques of statistical analysis". The NBS/NIST Standards Laboratory environment is conducive to the development and application of such tools and promotes the entire concept that the tools themselves should have a standardization component. Further, the 4 problem-areas that we have chosen to discuss here (univariate, interlab, comparative, and sensitivity/screening) are obviously important in science, engineering, and industrial pursuits everywhere, and thus the described tools will have applications well beyond NBS/NIST.

Conclusion

I thank the FTC/ASQ/ASA for the honor, privilege, and opportunity to provide this exposition of Jack Youden's remarkable career from the vantage point of a distant ancestor/alumnus at NBS/NIST/SEL. Further, I am most appreciative of ASQ Statistics Division (in the persons of Mindy Hotchkiss and Harish Jose) to graciously allow my Youden story to carry over to 2 editions of the ASQ Stat

Digest. Indeed, it would only be such a person with the capabilities, contributions, and impact of a Jack Youden that would necessitate the existence of such a 2-part series.

Jack Youden was the consummate problem-solver, consultant, and collaborator, who was a creative and seminal thinker in his development of relevant and straightforward design & analysis methodologies. He was a tireless NBS ambassador in his dissemination of solid, insightful methodology to the outside research community. He was a prolific, clear-minded writer, as well as a masterful oral communicator, committed to the enthusiastic evangelization of the virtues of rigorous statistical design and insightful statistical analysis. He set a standard for professional expertise and commitment that carries on to today.

We saw how his standards for excellence have served as the omnipresent backdrop for the NIST Statistics Division over the 5 decades of the post-Youden era. Though impossible to match, it has been possible to be strive for, and SEL's activities, awards, research efforts, publications, workshops, software, and stat methodology tools all reflect the legacy left behind by Jack Youden, the problem-solver. Even with 50 years of advancements in our computational and communication tools, the Youden legacy of excellence in consulting, communication, and methods development-all enhanced by his unbounded creativity and passion-still provides the framework for us to optimal and effective in assisting the scientists, engineers, researchers, and industrialists of the world.

We saw that Youden leveraged his expert knowledge of DEX & EDA to make for a better world-starting with better NBS standards. Even in his world of limited computer power, he showed us how to leverage DEX & EDA to be of maximal benefit to the researcher-in whatever discipline. We contended that Youden would approve of the 4 methodologies we recommended, and that there are abundant opportunities to apply standards for DEX and EDA development-not in the sense of limiting creativity, but in the sense of having such "standard" tools be routinely applied to "standard" problem categories. The net effect of this is to assure that underlying structure gets ferreted out and insight maximized. And analogously just as calibration standards are critical for increased scientific accuracy, we contended that statistical methodology standards are critical for increased insight.

Finally, we pose the hypothetical question that if Youden were alive today, what would he be doing? What problems would he be solving? What computer tools would he be using? What DEX/EDA methods would he be developing? We can all provide our own answers to these questions which may differ in detail, but for sure we know that his brainpower, servant-personality, commitment, passion, energy, versatility, creativity, and DEX/EDA talents would be just as effective today as it was 50+ years ago. Jack Youden would succeed as Jack Youden in any era.

References

- Box, G.E.P., Hunter, J.S., and Hunter, W.G. (1978) Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, Edition 1, p. 410, Wiley.
- Box, G.E.P., Hunter, J.S., and Hunter, W.G. (2005) Statistics for Experimenters: Design, Innovation, and Discovery, Edition 2, p. 272, Wiley.
- Croarkin, C, et al (2003) "NIST/SEMATECH e-Handbook of Statistical Methods", February 2003. https://www.itl.nist.gov/div898/handbook/
- Dataplot website: https://www.itl.nist.gov/div898/software/dataplot/
- e-Handbook website: https://www.itl.nist.gov/div898/handbook/
- Filliben, J.J. (1975), "The probability plot correlation coefficient test for normality", *Technometrics*, Volume 17, No. 1, February, pp. 111–117. DOI: 10.1080/00401706.1975.10489279
- Filliben, J.J. (1981) "DATAPLOT—an interactive high-level language for graphics, non-linear fitting, data analysis, and mathematics" ACM SIGGRAPH Computer

Graphics '81 Proceedings of the 8th Annual Conference on Computer Graphics and Interactive Techniques, Volume 15, Issue 3, August, pages 199–213, ACM New York, doi: 10.1145/800224.806807 http://dl.acm.org/citation.cfm?id=806807

- Filliben, J.J. (1984) "DATAPLOT—Introduction and Overview", NIST Special Publication 667, pdf version: https://www.itl.nist.gov/div898/software/dataplot/sp667.pdf

- Filliben, J.J., Cetinkunt, S., Yu, W.Y., and Donmez, A. (1993) "Exploratory Data Analysis Techniques as Applied to a High-Precision Turning Machine" in the book "Quality Through Engineering Design", edited by Kuo, W. and Pierson, M.M., Elsevier, New York, pp. 199–223. <block plot>

- Filliben, J.J. (2003) "Block Plot", Chapter 1.3.3.3 of "NIST/SEMATECH e-Handbook of Statistical Methods", February. https://www.itl.nist.gov/div898/handbook/eda/section3/blockplo.htm

- Filliben, J.J. et al (2003) "An EDA Approach to Experiment Design", Chapter 5.5.9 of "NIST/SEMATECH e-Handbook of Statistical Methods", February. https://www.itl.nist.gov/div898/handbook/pri/section5/pri59.htm

- Filliben, James J.(2020): Youden Address: The Role of DEX and EDA for Standards and the Role of Standards for DEX & EDA, Part 1", ASQ Statistics Digest, February 2020, Vol. 39, No. 1, p. 5–19. https://my.asq.org/communities/files/177/5133

- Fisher, R.A. (1925) "Statistical Methods for Research Workers". Oliver&Boyd (Edinburgh). ISBN 978-0-05-002170-5.

- Heckert, N.A., and Filliben, J.J. (2000), "Dataplot" http:s//www.itl.nist.gov/div898/software/dataplot/

- Ku, H.H., (1969) NBS Special Publication 300: Precision Measurement and Calibration-Statistical Concepts and Procedures https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nbsspecialpublication300v1.pdf

- Natrella, M.G. (1972) "Summary and Index for 'Statistical Design: A collection of the bimonthly articles by W. J. Youden that appeared in Industrial and Engineering Chemistry during the years 1954–1959'" *Journal of Quality Technology*, Vol. 4, No. 1, January 1972, pages 60–61. https://doi.org/10.1080/00224065.1972.11980515

- Scott, John Henry (2007), Accuracy issues in chemical and dimensional metrology in the SEM and TEM, Measurement Science & Technology, 18, July 20, 2007, pp. 2755–2761 DOI:10.1088/0957-0233/18/9/003

- Youden, W.J. (1951) *Statistical Methods for Chemists*, Wiley (126 pages).

- Youden, W.J. (1960) *Statistical Design*, A 5-year (1954–1959) bimonthly collection of 36 articles by Jack Youden, reprinted by the American Chemical Society from IEC: Industrial & Engineering Chemistry (72 pages).

- Youden, W.J. (1962) *Experimentation and Measurement*. Vistas of Science Series, National Science Teachers Association, Washington, DC (128 pages).

- Youden, W.J. (1967) *Statistical Techniques for Collaborative Tests*, The Association of Official Analytical Chemists, Box 540, Benjamine Franklin Station, Washington, DC (64 pages).

- Youden, W.J. (1974) *Risk, Choice, and Prediction: An Introduction to Experimentation, Duxbury Press (81 pages).*

Statistical Process Control

Donald J. Wheeler, Statistical Process Controls, Inc.

Covid-19 Data and Process Behavior Charts

Every day we are told how many new confirmed cases of Covid-19 were reported and how many people have died of this disease. How do we use these numbers? Routinely newspapers get it wrong. (A recent headline reported the number as "soaring" when in fact they were reporting the smallest daily change in the past month.) So how can we overcome such a total misunderstanding of the data? Only by putting the data in context.

Table 1 shows the daily number of new confirmed cases of Covid-19 in the U.S. These are the values posted by the European CDC at noon London time, thus they are slightly smaller than values that are reported later each day. Typically, these daily numbers are graphed as a bar chart. Figure 1 shows this bar chart for the data in Table 1 up to 4/14.

Table 1: Daily Number of New Confirmed Covid-19 Cases in U.S.

3/3	14	3/25	8,789	4/16	30,148
3/4	22	3/26	13,963	4/17	31,667
3/5	34	3/27	16,797	4/18	30,833
3/6	74	3/28	18,695	4/19	32,922
3/7	105	3/29	19,979	4/20	24,601
3/8	95	3/30	18,360	4/21	28,065
3/9	121	3/31	21,595	4/22	37,289
3/10	200	4/1	24,998	4/23	17,588
3/11	271	4/2	27,103	4/24	26,543
3/12	287	4/3	28,819	4/25	21,352
3/13	351	4/4	32,425	4/26	48,529
3/14	511	4/5	34,272	4/27	26,857
3/15	777	4/6	25,398	4/28	22,541
3/16	823	4/7	30,561	4/29	24,132
3/17	887	4/8	30,613	4/30	27,326
3/18	1,766	4/9	33,323	5/1	29,917
3/19	2,988	4/10	33,901	5/2	33,955
3/20	4,835	4/11	35,527	5/3	29,288
3/21	5,374	4/12	28,391	5/4	24,972
3/22	7,123	4/13	27,620	5/5	22,593
3/23	8,459	4/14	25,023	5/6	23,841
3/24	11,236	4/15	26,922	5/7	24,128

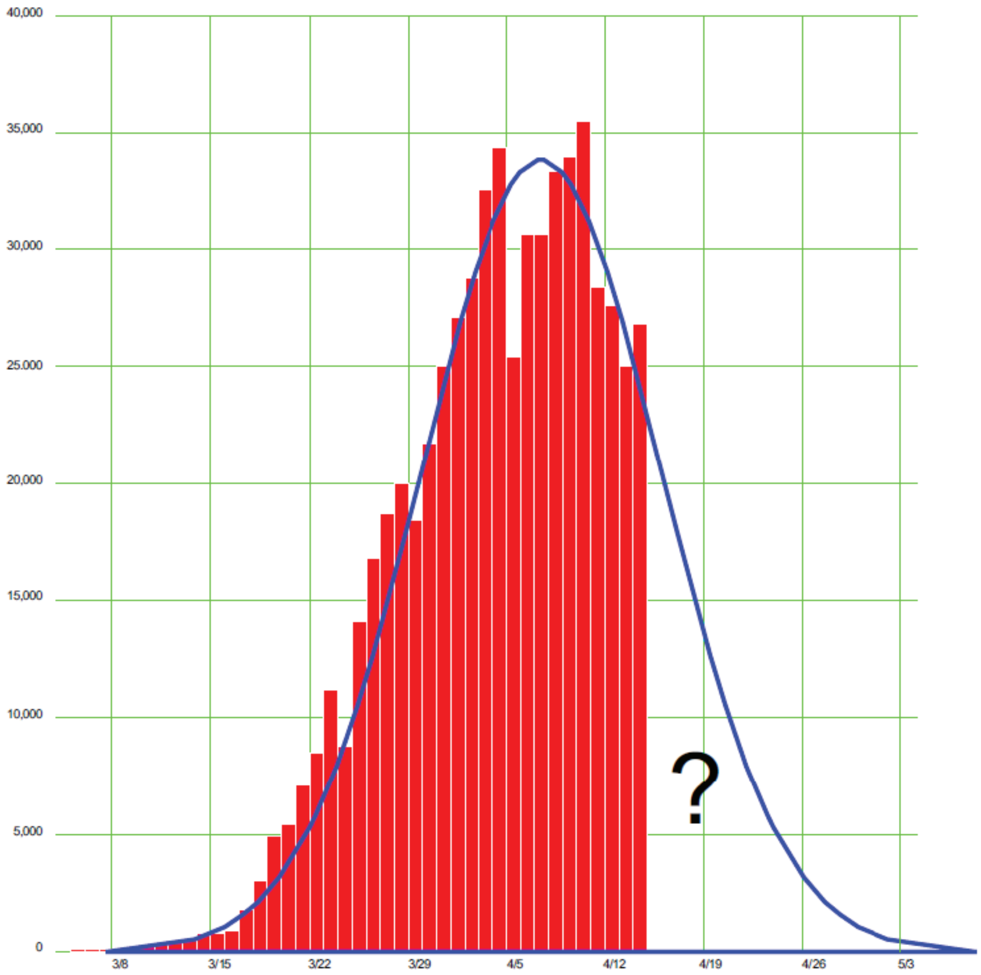


Figure 1: Bar Chart for Daily Number of New Confirmed Covid-19 Cases in U.S.

As the values climb up to some potential peak this bar chart often makes people think of a probability distribution. We look for "the peak," and want to know when we have gotten "over the hump." Confusing this bar graph with a histogram leads to all sorts of crazy ideas. One correspondent suggested that we might compute a "z-score" using each country's version of Figure 1. This would allow us to quantify how far along the "curve" each country was located. (Clearly Figure 1 shows that the U.S. has a z-score that is approaching +1, and the epidemic will be over by early May!)

Unfortunately, the data in Table 1 do not represent a probability distribution. They are a time series. And according to William Playfair, the man who invented the bar chart, *we should not place time series data on a bar chart.* The reason being that the bar chart is for comparing amounts. A bar chart draws our attention to the vertical heights of the bars. But with a time series our mind wants to know how things are changing over time. To get a graph that will draw our eyes in the direction that our mind wants to go, to see how the series changes over time, we need to use a running record rather than a bar chart.

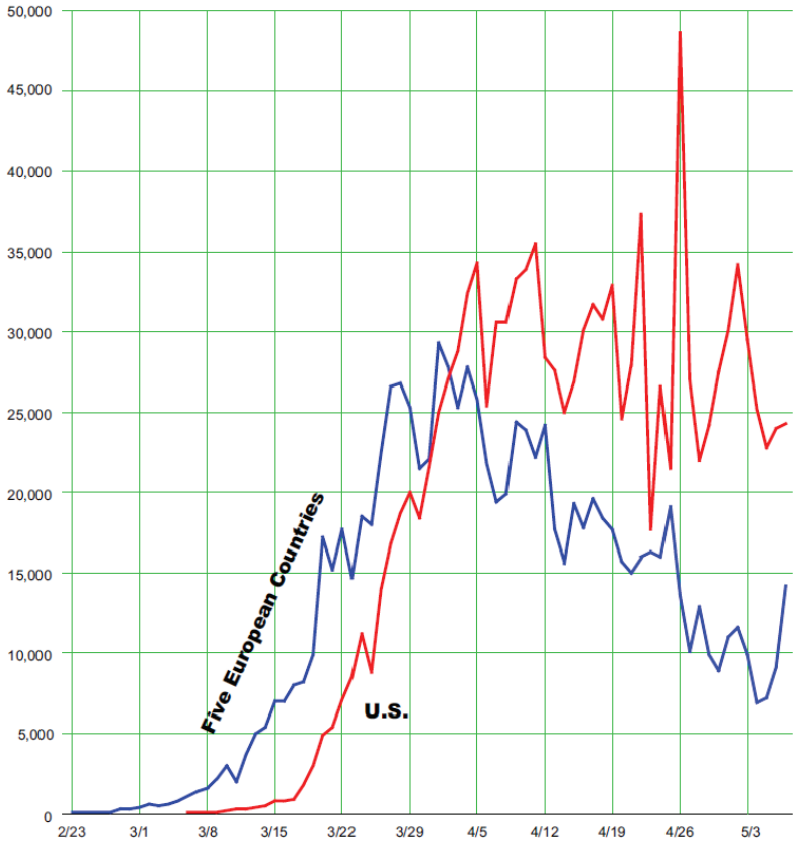


Figure 2: The Daily Number of New Confirmed Covid-19 Cases in U.S.

By getting rid of the bars, we can even compare running records. Figure 2 shows the daily numbers of new Covid-19 cases. The red line comes from Table 1. The blue line is the combined daily counts for the U.K., Germany, France, Spain, and Italy. Since these five countries combined have essentially the same population as the U.S., the time series in Figure 2 make a reasonable comparison. In addition, as of May 7, these six countries represent over 58% of the World's cases of Covid-19, and 71% of the World's Covid-19 deaths, even though they only contain 9% of the World's population.

Figure 2 shows that although the Covid-19 pandemic began about 10 days earlier in Europe than in the U.S., the U.S. had caught up within a month. In both cases the number of new cases climbed quickly at first. Then, as interventions came into effect, the numbers plateaued. While it is very hard to make any case for a decline in the U.S. numbers, there does appear to be a slow decline in Europe over the past few weeks.

Notice that no "analysis" was required to make sense of Figure 2. When we know that things are changing, as they always are in any epidemic, the running record becomes self-interpreting. While the purpose of analysis is insight, the best analysis is always the simplest analysis that provides the needed insight. After all, we have to share our discoveries with others, and it is always easier to explain a simple analysis than a complicated one. And what can be simpler than letting the data speak for themselves?

NONSENSE HAPPENS

Nevertheless, people all over the World are trying to use statistical axes to "analyze" the Covid data. After all, we can't possibly consume "raw" data. We have to process it first to make it unintelligible.

One of these statistical axes is the process behavior chart. Process behavior charts are incredibly useful and versatile. They allow us to filter out the noise so we can pay attention to the signals. They allow us to identify those points in time where a change occurs in an otherwise steady-state system. As one colleague of mine quipped, they ought to be called "has-a-change-occurred" charts.

But, as we have seen, an epidemic is anything but a steady-state system. It grows, evolves, and eventually declines. Here we do not need to ask the question "Has a change occurred?" because we know that the epidemic is constantly changing.

No, the question here is "How is the epidemic currently changing?" And the only useful answers for this question depend on estimates of the growth rate. How to estimate this growth rate will be explained later.

IGNORING THE NATURE OF THE DATA

Nevertheless, some use the Covid numbers to compute related values like the number of people recovered and the number of people ill, and then put these numbers on process behavior charts to ask if these values are changing. The problem here begins with the computation. The number of confirmed cases and the number of people who have recovered are nothing more than *lower bounds* on these categories. They are "confirmed" cases and "confirmed" recoveries.

When those who were repatriated from China were tested they found that 40% to 50% of the Covid-19 infections were either asymptomatic or so mild as to have been missed without the tests. Thus, any values computed from the confirmed numbers are going to be incomplete, and any analysis of such values will invariably be very insensitive.

Others seek to use process behavior charts to identify change points in the progress of the epidemic. However, given the nature of an epidemic, change is everywhere, and it is the longer-term trends that matter more than the daily values.

Still others wait for the plateau of Figure 2 and then try to use process behavior charts to separate days with 20,000, 30,000, and 40,000 new cases into categories of "common cause days" and "special cause days." Since we are dealing with a natural phenomenon, where we have no real process inputs, these categories have no real meaning here. (Hint: all of these counts are undesirable regardless of how the days are categorized.)

SEVEN STEPS TO NOWHERE

And then there are those who get so completely carried away with using their analysis techniques that they lose sight of what the data represent. Several of these have been teaching a seven step analysis: for Covid-19 data.

(1) Begin by assuming the daily counts are modeled by a Poisson distribution and use c-charts to identify when the epidemic starts to "grow."

(2) When we have a signal of growth, transform the counts by taking logarithms.

(3) Obtain a regression equation using the logarithms of the daily counts in order to estimate the growth rate.

(4) Create an *XmR* chart with sloping limits around the regression equation.

(5) Re-transform the limits in order to plot them on top of the running record of Figure 2.

(6) Continue to plot points on Figure 2 until you get a point outside the exponentially increasing limits (on the right-hand side).

(7) On the date corresponding to this point outside the limits declare the epidemic to have "peaked." (This latter point will generally be found a few days after the day where the curve in Figure 3 below begins to flatten out.)

At this point no insight has been created by all of this complexity and analysis. Everything "found" by this seven-step analysis, the initial growth, the growth rate, and the beginning of the flattening of the curve, is already made visible in the simple running records of Figures 2 and 3. No value is added by the seven step analysis. It simply uses a lot of computing power in order to decorate Figure 2 with what Edward Tufte calls "non-data ink", or more concisely, "chartjunk." And of course, they will be glad to sell you some software to create all this chartjunk.

BUT ARE WE DOING BETTER OR WORSE?

This is the question everyone has on their mind when they hear the daily Covid numbers. But the large swings in Figure 2 prevent us from seeing the big picture and answering this question. Some days look promising, and others look bad. Before we can see the big picture two things have to happen. First, we have put the number of new cases in context by using the total number of cases to date, and second, we have to plot these totals on a semi-log plot.

When we combine the daily new cases into the total to date we dampen out the day-to-day swings by the ballast of the history of the epidemic. So using the totals is a smoothing technique.

When we plot the totals-to-date on a semi-log plot we make the *growth rate of the epidemic* visible on the graph. Since the growth rate directly answers the question of "Are we doing better or worse?" it is a key descriptor for any epidemic.

When we use linear coordinate axes we make *amounts* visible and comparable—a one unit change means the same thing everywhere on the graph. But when we use a semi-log plot we make *growth rates* visible by turning fixed rate growth (exponential growth) into straight lines. If the rate of growth remains the same the growth curve will be a straight line on a semi-log plot. Moreover, parallel line segments will always represent the same rate of growth regardless of where they appear on a semi-log plot.

So, when we plot the total number of Covid-19 cases to date on a semi-log plot we know that *any persistent deviation from a straight line will represent a change in the growth rate for the pandemic*. It really is that easy.

Table 2 shows the total number of confirmed cases of Covid-19 in the U.S. They come from the same source as Table 1. Figure 3 shows the data from Table 2 plotted along with the corresponding combined totals for the U.K., Germany, France, Spain, and Italy. In addition, Figure 3 includes the daily totals of confirmed cases for South Korea, Norway, and Australia.

Table 2: Cumulative Number of Confirmed Covid-19 Cases in U.S.

3/3	103	3/25	55,231	4/16	639,664
3/4	125	3/26	69,194	4/17	671,331
3/5	159	3/27	85,991	4/18	702,164
3/6	233	3/28	104,686	4/19	735,086
3/7	338	3/29	124,665	4/20	759,687
3/8	433	3/30	143,025	4/21	787,752
3/9	554	3/31	164,620	4/22	825,041
3/10	754	4/1	189,618	4/23	842,629
3/11	1,025	4/2	216,721	4/24	869,172
3/12	1,312	4/3	245,540	4/25	890,524
3/13	1,663	4/4	277,965	4/26	939,053
3/14	2,174	4/5	312,237	4/27	965,910
3/15	2,950	4/6	337,635	4/28	988,451
3/16	3,774	4/7	368,196	4/29	1,012,583
3/17	4,661	4/8	398,809	4/30	1,039,909
3/18	6,427	4/9	432,132	5/1	1,069,826
3/19	9,415	4/10	466,033	5/2	1,103,781
3/20	14,250	4/11	501,560	5/3	1,133,069
3/21	19,624	4/12	529,951	5/4	1,158,041
3/22	26,747	4/13	557,571	5/5	1,180,634
3/23	35,206	4/14	582,594	5/6	1,204,475
3/24	46,442	4/15	609,516	5/7	1,228,603

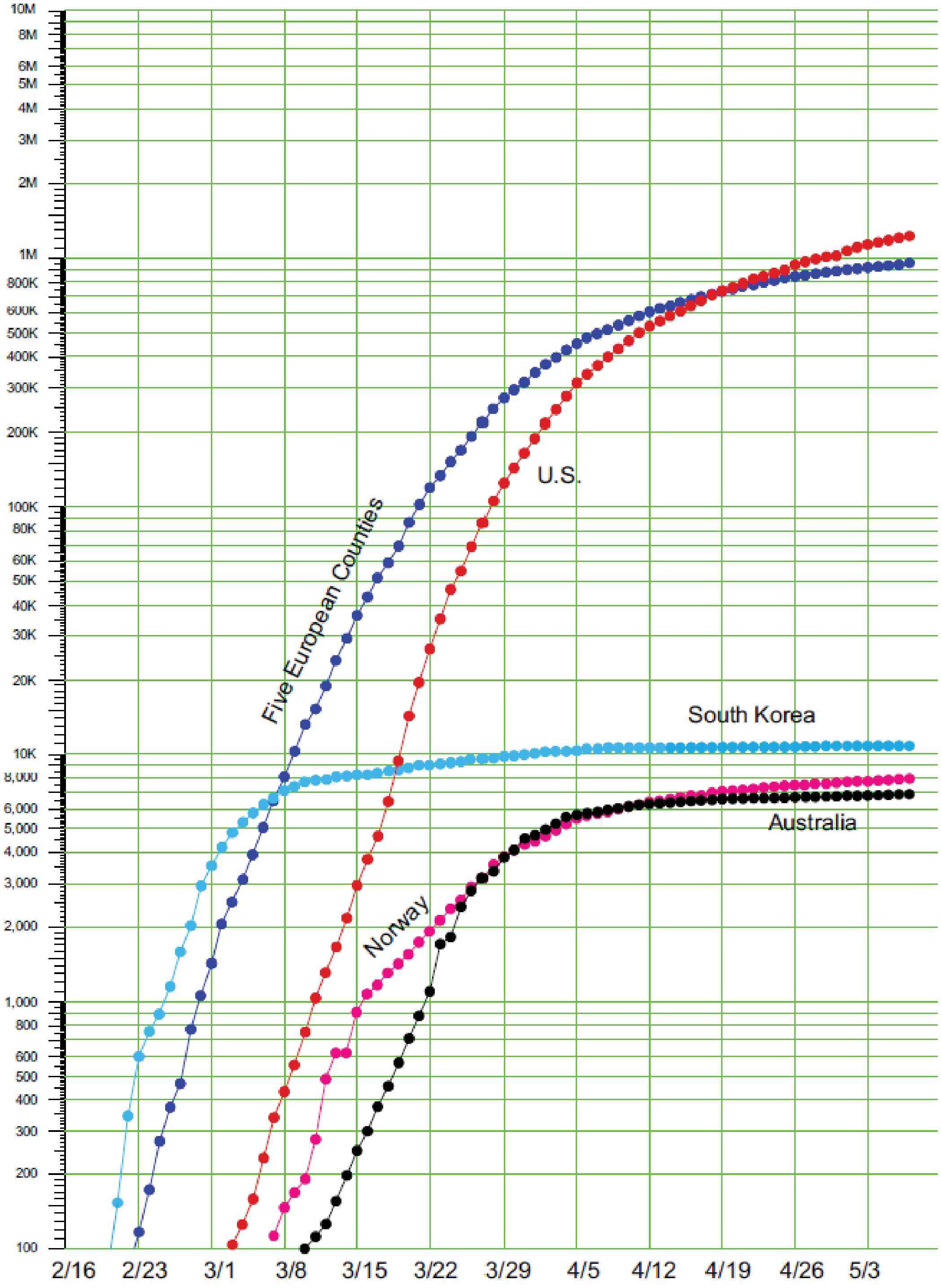


Figure 3: Total Confirmed Cases in the U.S., Five European Countries, South Korea, Norway, and Australia

Any curve on a semi-log plot represents a changing growth rate. As these curves flatten out over time the growth rate for the Covid-19 pandemic is slowing. Where they are steeper the growth rate is greater. Once more we have a self-interpreting, easy to understand graph of the raw data.

In Figure 3 we see that everyone has flattened their curves. But while everyone started out with very similar growth rates, the U.S. has not flattened its curve as much as the others. Consequently the U.S. total count passed that of the five European countries on April 19. On the other hand, South Korea, Norway, and Australia illustrate what modern democracies can accomplish in terms of flattening the curve. All of this is immediately clear from Figure 3.

Moreover, Figure 3 gives a simple way to make reasonable forecasts. When we use the current slope of a curve and extend a straight line we are simply projecting what can be expected without any further changes in the growth rate.

QUANTIFYING THE CURVES

One of the ways to characterize the rate of growth is the number of days needed for the total count to double. This doubling time can easily be obtained either from the graph or from the table. For example, from Table 2, on March 11 there were 1025 confirmed Covid-19 cases in the U.S. On March 14 there were 2174. Thus, in mid-March, the number of Covid cases was doubling about every three days. This estimate applies everywhere along the reasonably straight portion of the U.S. curve between 3/4 and 3/17.

From 4/4 to 4/13, the doubling time increased to about nine days as may be seen by comparing 277,945 with 557,571 in the table.

How many days did it take to go from one-half million cases to one million cases in Table 2? The problem with this eyeball approach is that with the larger values the curve does not always remain straight for a long enough period, so we need a way to actually compute a growth rate.

To use the data to estimate the growth-rate of an epidemic we choose a baseline of several days and compute an average daily gain. To compute an average daily gain for the period of 3/11 to 3/14 we begin with the four sequential values:

3/11	1,025
3/12	1,312
3/13	1,663
3/14	2,174

Divide each day’s value by the value for the previous day:

$2174/1663 = 1.307$

$1663/1312 = 1.268$

$1312/1025 = 1.280$

Average these day-to-day gains to get:

$$\textit{Average Daily Gain} = 1.285$$

So during this period the epidemic was growing at an estimated rate of 28.5% per day. This growth rate can be converted into a doubling time by the formula:

$$\textit{Doubling Time in Days} = \frac{\log[2.000]}{\log[\textit{Average Daily Gain}]}$$

Since we are working with a ratio of logarithms it does not matter which base we use for these logarithms (as long as we use the same base for both). Here our average daily gain of 1.285 yields a doubling time of 2.76 days, which agrees with the simple estimate of about three days found above.

During the last seven days shown in Table 2 the U.S. growth rate has averaged 2.3%. This corresponds to a doubling time of 30 days. This doubling time gives us a benchmark to use in judging how we are doing in the future. Since we are currently at 1.23 million, a doubling time of 30 days suggests that by the first week in June we might have as many as 2.4 million cases. If we do better than this, we will see it on the graph. If not, then this is what we should expect.

South Korea, Norway, and Australia show what can be accomplished in terms of flattening the curve. They have growth rates of 0.1%, 0.6%, and 0.3% per day respectively. The five European countries combined have a growth rate of 1.2%, and the U.S. is clearly in last place in this race.

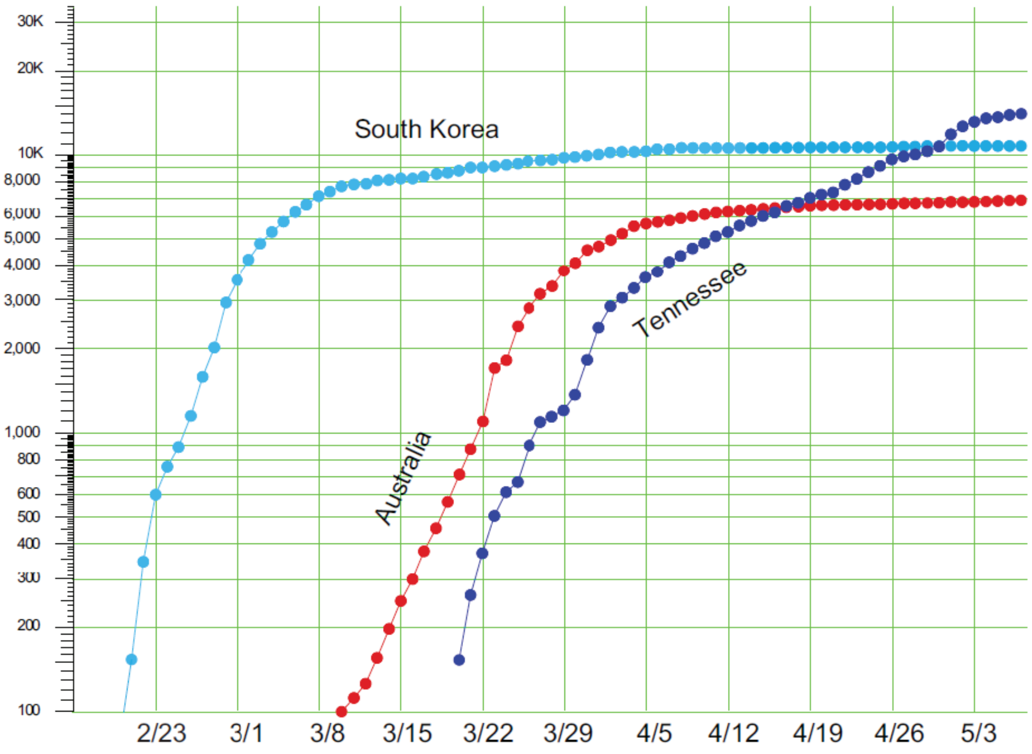


Figure 4: Total Confirmed Covid-19 Cases in South Korea, Australia, and Tennessee

Figure 4 shows the daily totals for my home state of Tennessee along with those for South Korea and Australia. Tennessee has 6.7 million residents. Australia has 25 million residents. Tennessee's count of confirmed Covid-19 cases passed that of Australia on April 17. South Korea has 51.6 million residents, yet Tennessee's count passed Korea's count on May 1.

SUMMARY

The best analysis is the simplest analysis that provides the needed insight. With data from an epidemic there is no question of whether a change has occurred. Change is everywhere. The question is whether we are getting better or worse.

So while the process behavior chart may be the Swiss army knife of statistical techniques, there are times when we need to leave the knife in our pocket, plot the data, and then listen to them as they tell their story.

Hypothesis Testing

Jim Frost

Understanding Significance Levels

Significance levels in statistics are a crucial component of hypothesis testing. However, unlike other values in your statistical output, the significance level is not something that statistical software calculates. Instead, you choose the significance level. Have you ever wondered why?

In this column, I'll explain the significance level conceptually, why you choose its value, and how to choose a good value. Statisticians also refer to the significance level as alpha (α).

First, it's crucial to remember that hypothesis tests are inferential procedures. These tests determine whether your sample evidence is strong enough to suggest that an effect exists in an entire population. Suppose you're comparing the means of two groups. Your sample data show that there is a difference between those means. Does the sample difference represent a difference between the two populations? Or, is that difference likely due to random sampling error? That's where hypothesis tests come in!

Your sample data provide evidence for an effect. The significance level is a measure of how strong the sample evidence must be before determining the results are statistically significant. Because we're talking about evidence, let's look at a courtroom analogy.

Evidentiary Standards in the Courtroom

Criminal cases and civil cases vary greatly, but they both require a minimum amount of evidence to convince a judge or jury to prove a claim against the defendant. Prosecutors in criminal cases must prove the defendant is guilty "beyond a reasonable doubt," whereas plaintiffs in a civil case must present a "preponderance of the evidence." These terms are evidentiary standards that reflect the amount of evidence that civil and criminal cases require.

For civil cases, most scholars define a preponderance of evidence as meaning that at least 51% of the evidence shown supports the plaintiff's claim. However, criminal cases are more severe and require stronger evidence, which must go beyond a reasonable doubt. Most scholars define that evidentiary standard as being 90%, 95%, or even 99% sure that the defendant is guilty.

In statistics, the significance level is the evidentiary standard. For researchers to successfully make the case that the effect exists in the population, the sample must contain a sufficient amount of evidence.

In court cases, you have evidentiary standards because you don't want to convict innocent people.

In hypothesis tests, we have the significance level because we don't want to claim that an effect or relationship exists when it does not exist.

Significance Levels as an Evidentiary Standard

In statistics, the significance level defines the strength of evidence in probabilistic terms. Specifically, alpha represents the probability that tests will produce statistically significant results when the null hypothesis is correct. Rejecting a true null hypothesis is a type I error. And, the significance level equals the type I error rate. You can think of this error rate as the probability of a false positive. The test results lead you to believe that an effect exists when it actually does not exist.

Obviously, when the null hypothesis is correct, we want a low probability that hypothesis tests will produce statistically significant results. For example, if alpha is 0.05, your analysis has a 5% chance of producing a significant result when the null hypothesis is correct.

Just as the evidentiary standard varies by the type of court case, you can set the significance level for a hypothesis test depending on the consequences of a false positive. By changing alpha, you increase or decrease the amount of evidence you require in the sample to conclude that the effect exists in the population.

Changing Significance Levels

Because 0.05 is the standard alpha, we'll start by adjusting away from that value. Typically, you'll need a good reason to change the significance level to something other than 0.05. Also note the inverse relationship between alpha and amount of required evidence. For instance, increasing the significance level from 0.05 to 0.10 lowers the evidentiary standard. Conversely, decreasing it from 0.05 to 0.01 increases the standard. Let's look at why you would consider changing alpha and how it affects your hypothesis test.

Increasing the Significance Level

Imagine you're testing the strength of party balloons. You'll use the test results to determine which brand of balloons to buy. A false positive here leads you to buy balloons that are not stronger. The drawbacks of a false positive are very low. Consequently, you could consider lessening the amount of evidence required by changing the significance level to 0.10. Because this change decreases the amount of required evidence, it makes your test more sensitive to detecting differences, but it also increases the chance of a false positive from 5% to 10%.

Decreasing the Significance Level

Conversely, imagine you're testing the strength of fabric for hot air balloons. A false positive here is very risky because lives are on the line! You want to be very confident that the material from one manufacturer is stronger than the other. In this case, you should increase the amount of evidence required by changing alpha to 0.01. Because this change increases the amount of required evidence, it makes your test less sensitive to detecting differences, but it decreases the chance of a false positive from 5% to 1%.

It's all about the tradeoff between sensitivity and false positives!

In conclusion, a significance level of 0.05 is the most common. However, it's the analyst's responsibility to determine how much evidence to require for concluding that an effect exists. How problematic is a false positive? There is no single correct answer for all circumstances. Consequently, you need to choose the significance level!

While the significance level indicates the amount of evidence that you require, the p-value represents the strength of the evidence that exists in your sample. When your p-value is less than or equal to the significance level, the strength of the sample evidence meets or exceeds your evidentiary standard for rejecting the null hypothesis and concluding that the effect exists.

Risk and Uncertainty

S. Luko, Collins Aerospace

Interval Estimation

In this edition of "Risk and Uncertainty" we'll discuss the important topic of interval estimation and its relation to uncertainty. The interval estimation by definition has to do with uncertainty. There are so many different cases for statistical intervals that we can barely do it justice here. The excellent Wiley text, [1], by Hahn and Meeker, Statistical Intervals, as well as the many papers by Hahn and others are the go-to sources for many of the details of this topic. Here we try and bring attention to a sample of the most popular cases.

Without doubt, the interval most people are familiar with is the confidence interval. To review, a confidence interval applies to an unknown parameter of a parametric distribution. This includes continuous and discrete cases. For continuous variables familiar examples include the normal, and lognormal distributions; for discrete cases, familiar examples include the binomial and Poisson distributions. Many other cases might be added. Confidence intervals were developed by the Polish mathematician Jerzy Neyman and presented in a famous 1937 paper, *Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability* [2].

What can we say about confidence intervals? All of the main pantheon of applications have well known formulas—means and variances for a normal distribution, the binomial event probability p, the Poisson and exponential rate constant λ. There are also the well-known cases of comparing two means, variances, event probabilities and rates. The one important point to make about confidence intervals is that, as a final result, we can't make any probability statement about the interval. For example, suppose we have constructed a 95% confidence interval for a mean from a normal distribution and the final result is: $150 \le \mu \le 165$. We can't say that $P(150 \le \mu \le 165) = 0.95$. That's because there is no longer a random variable within the parentheses that could vary according to a probability distribution. The mean is either in the interval or it is not. Let us now look at the random variable theoretic rendition for this interval.

$$\bar{x} - \frac{t_{\alpha/2}S}{\sqrt{n}} \le \mu \le \bar{x} + \frac{t_{1-\alpha/2}S}{\sqrt{n}} \tag{1a}$$

Where the sample mean and standard deviation are the random variables $\bar{x}$ and s, n is the sample size, and $t_{\alpha/2}$ and $t_{1-\alpha/2}$ are the appropriate quantiles from a t distribution with $n-1$ degrees of freedom to make the probability of this statement equal to $1-\alpha$. In this rendition it is perfectly legitimate to state that:

$$P\left(\bar{x} - \frac{t_{\alpha/2}S}{\sqrt{n}} \le \mu \le \bar{x} + \frac{t_{1-\alpha/2}S}{\sqrt{n}}\right) = 1-\alpha \tag{1b}$$

The reason is that there is no data being used, just the two random variables $\bar{x}$ and s. The quantity μ is the fixed unknown mean of the normal we are sampling from. If we were to execute this interval many thousands of times, each time sampling from the same normal distribution and with differing data each time, we would find that approximately $100(1-\alpha)$% of the time the constructed interval would contain or "capture" the true value μ. In practice we only have one example of this and once constructed we only have numbers for each limit—not random quantities anymore. Since, it is the hypothetical process we are working with (i.e. one example from many possible), we use the term "confidence" in place of probability. Some statisticians use the notation $C(L \le \mu \le U) = 100(1-\alpha)$%, where "C" signifies that the constructed interval is a confidence interval, and L and U are the limits of the confidence interval determined using (1a) with the available data.

There are several ways to show how this works graphically using simulation. Here is a short favorite example. Suppose we sample from a normal distribution with $\mu = 1000$ and $\sigma = 10$. Use a sample size of $n = 10$ and repeat the sampling 10,000 times saving

$\bar{x}$ and s each time. For each case, construct 95% confidence limits (L and U) for μ using (1a). Use t = 2.26216 (the upper 97.5% quantile from a t distribution with 9 degrees of freedom). Save the statistics L and U. Plot L against U in a scatter plot. This shows a "cloud" of 10,000 possible (L, U) pairs for this sampling case. An example of this result is shown in Figure 1. In the figure draw two lines cutting the plot into four quadrants. The lines are drawn at the true value μ = 1000 for both axis. Then the lower right hand quadrant is the case where L ≤ 1000 and U ≥ 1000, meaning that L ≤ μ ≤ U for this case where we have used μ = 1000. In this example, we find that 95.04% of the (L, U) points lie in the lower right hand quadrant so that approximately 95% of the time these pairs are correct in that μ = 1000 is contained within the interval. The remaining 5% of cases are shown in the upper right and lower left quadrants where μ = 1000 lies outside the (L, U) interval. In practice we only have one case and that is like sampling a random point from this cloud.

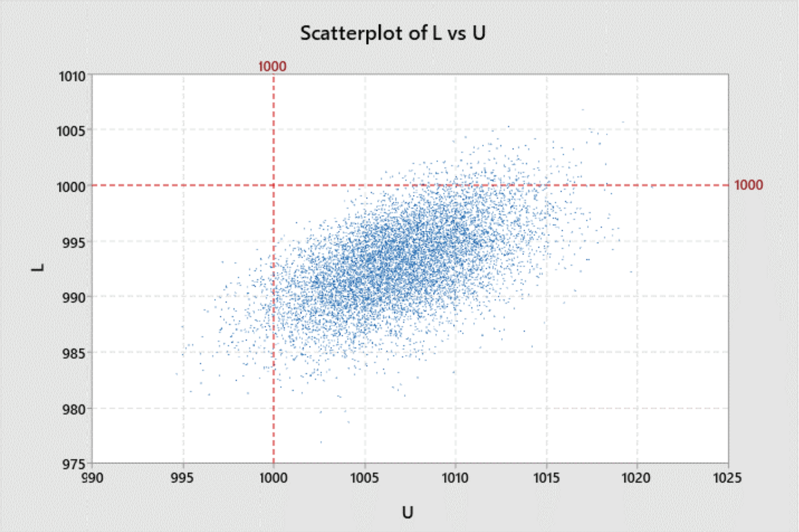


Figure 1: Example of Confidence Limits L and U sampled from a normal distribution with μ = 1000 and σ = 10 and n = 10; Number of iterations 10,000

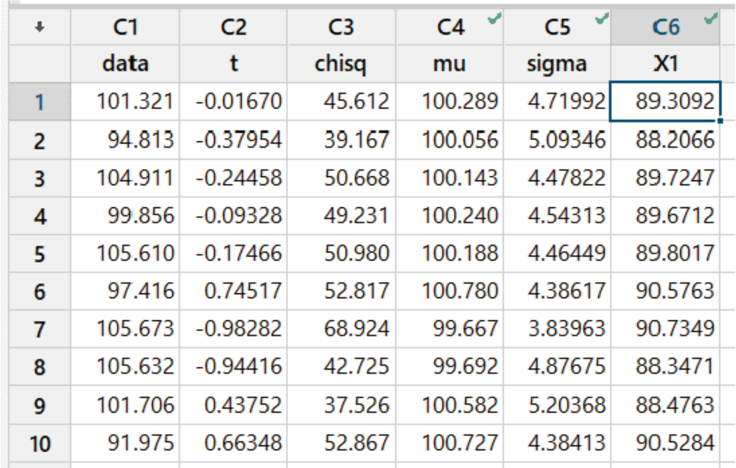
The constructed limits L and U and the width, w = U-L are also random variables in this setting. It is interesting to note that the width, w, of the confidence interval can be quite variable. In this example the empirical 1st and 99th percentiles of w were 6.872 and 22.402 respectively. Of course there are one sided variations of the confidence interval where the intervals are of the form (L, +∞) or (–∞, U).

A handy method that uses Monte Carlo simulation can be used for numerous cases where normal distributions are used. For example, in creating confidence intervals for percentiles, for mean differences and for the quality metrics Cpk or Ppk. The method starts with a set of data assumed to come from a process in a state of statistical control and normally distributed with unknown parameters μ and σ. Then we seek possible (μ, σ) pairs that could have generated the data. This is the so called fiducial confidence interval, [3], construction technique. When the data, x, are normally distributed and $\bar{x}$ and s are the sample mean and standard deviation in a sample of n observations, the following two well-known results provide the way forward with this.

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}} \qquad (2a)$$

$$y = \frac{(n-1)S^2}{\sigma^2} \qquad (2b)$$

The variable, t, has student's t distribution with n – 1 degrees of freedom; the variable y has a chi-square distribution with n – 1 degrees of freedom. t and y are independent. Start by generating 100,000 t's and y's in a program such as Minitab. Save these to two columns. Set t and y equal to the expressions (2a) and (2b) solving for μ and σ respectively. Save the (μ, σ) pairs to two columns. These last two columns constitute the (μ, σ) plausible values for the parameters that could have generated the data we see. The (μ, σ) pairs are then used to calculate any number of metrics for the process. From the distribution of such metrics the confidence interval can be calculated. Suppose we are interested in the 1st percentile of the distribution of individuals from the process. Calculate $\mu - 2.326\sigma$ for each case. There will result 100,000 cases from which the confidence interval for the 1st percentile may be studied. Figure 2 shows a screen shot of the data organization for this execution in Minitab. The data constitutes a random sample of $n = 50$ where $\bar{x} = 100.300$ and $s = 4.5538$ are the sample statistics.



↓	C1	C2	C3	C4	C5	C6
	data	t	chisq	mu	sigma	X1
1	101.321	-0.01670	45.612	100.289	4.71992	89.3092
2	94.813	-0.37954	39.167	100.056	5.09346	88.2066
3	104.911	-0.24458	50.668	100.143	4.47822	89.7247
4	99.856	-0.09328	49.231	100.240	4.54313	89.6712
5	105.610	-0.17466	50.980	100.188	4.46449	89.8017
6	97.416	0.74517	52.817	100.780	4.38617	90.5763
7	105.673	-0.98282	68.924	99.667	3.83963	90.7349
8	105.632	-0.94416	42.725	99.692	4.87675	88.3471
9	101.706	0.43752	37.526	100.582	5.20368	88.4763
10	91.975	0.66348	52.867	100.727	4.38413	90.5284

Figure 2: Screen Shot for the Fiducial Monte Carlo Example—From Minitab

From the data window, C1 contains the initial date ($n = 50$); C2 and C3 are the random and chi-square values; C4 and C5 are the recovered μ and σ pairs; and C6 is the calculated 1st percentile using μ and σ. Figure 3 shows the distribution of X1 resulting from this execution, Table 1 shows the empirical percentiles of the distribution of X1 where we show the confidence 95% interval limits in red. For this example, the interval is 86.805 to 99.887.

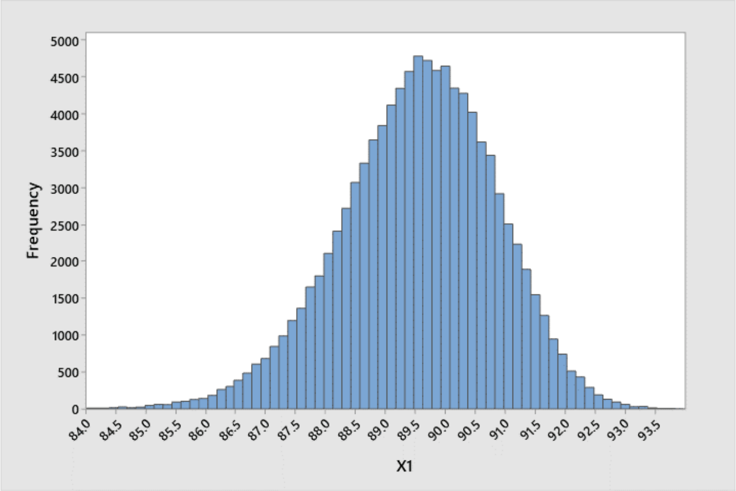
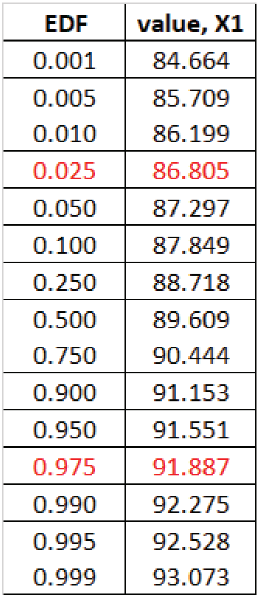


Figure 3: Empirical Results, 1st Percentile, Fiducial Monte Carlo, 100,000 Cases

Table 1: Empirical Frequency Distribution of Figure 3 Data.



EDF	value, X1
0.001	84.664
0.005	85.709
0.010	86.199
0.025	86.805
0.050	87.297
0.100	87.849
0.250	88.718
0.500	89.609
0.750	90.444
0.900	91.153
0.950	91.551
0.975	91.887
0.990	92.275
0.995	92.528
0.999	93.073

It is an easy exercise to develop the quality metric Ppk for this data. The specification requirement is bi-lateral, 100 ± 18 or 82 to 118. We add two columns using a formula, "mu-82" and "118-mu", where "mu" is the value of μ in column C4. A third additional column is added with a formula that calculates the minimum of (mu-82, mu-88) divided by "3σ" where "σ" is the column C5 entry. The final result is 100,000 Ppk values that might have given rise to the data. Figure 4 shows this result with the 10th percentile indicated. That number is about 1.1; Table 2 is the associated empirical frequency distribution.

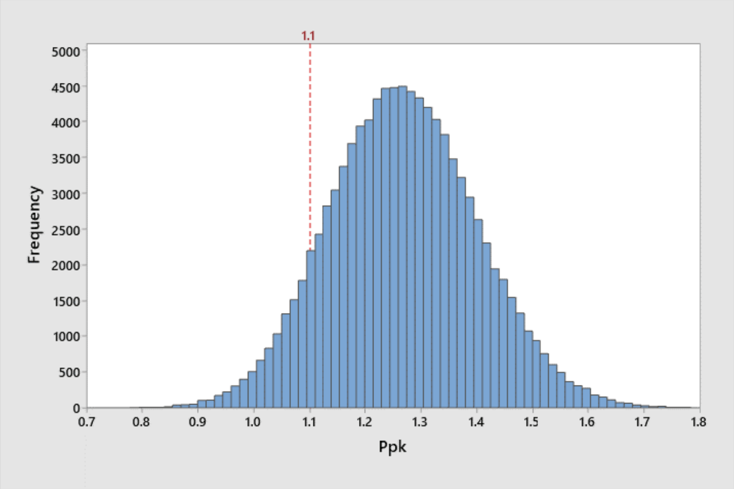
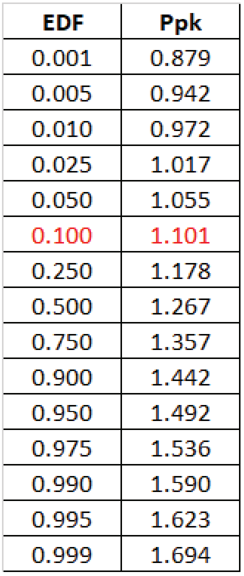


Figure 4: Empirical Results, Ppk, Fiducial Monte Carlo, 100,000 Cases

Table 2: Empirical Frequency Distribution of Figure 4 Data



EDF	Ppk
0.001	0.879
0.005	0.942
0.010	0.972
0.025	1.017
0.050	1.055
0.100	1.101
0.250	1.178
0.500	1.267
0.750	1.357
0.900	1.442
0.950	1.492
0.975	1.536
0.990	1.590
0.995	1.623
0.999	1.694

This data then shows the lower 90% confidence bound on the true Ppk to be about 1.1. Still other metrics may be studied using this technique. For example, for each μ, σ pair we could generate another sample and calculate various statistics such as the sample min or max.

Prediction Intervals

The prediction interval is known to almost anyone who has taken a first statistics course where simple linear regression is part of the course. What is surprising is the limited status in basic courses of the simple prediction interval when using the normal distribution. Its derivation is so similar to the confidence interval that it is worth summarizing it here. First, the prediction interval for a single future observation is an interval that would contain a single future observation with some confidence C. This can be important in a single shipment of a product that only happens occasionally, but the characteristic of concern is known from prior data to be normally distributed.

Suppose X is normal with some mean and variance, μ and σ^2. Let $\bar{x}$ and s stand for the sample mean and standard deviation from a sample of n observations. Let y be a future value, unrealized, from this normal distribution. The combination $y-\bar{x}$ is also normally distributed with mean 0; further, y and $\bar{x}$ are independent and the variance of $y-\bar{x}$ is then $\sigma^2(1 + 1/n)$. Then $Z=(y-\bar{x})/\left\{\sigma\sqrt{1+1/n}\right\}$ is standard normal and we have immediately that:

$$P\left(\bar{x}-Z_{\alpha/2}\sigma\sqrt{1+1/n}\le y\le\bar{x}-Z_{1-\alpha/2}\sigma\sqrt{1+1/n}\right)=1-\alpha \tag{3}$$

This is the known σ case. Notice that there are two random variables in (3) - $\bar{x}$ and y. We'll use the previous example where the population being sampled was Normal with $\mu = 1000$ and $\sigma = 10$ to study the behavior. Select 10,000 samples of $n = 10$, calculate and save $\bar{x}$ for each case. Construct intervals (L, U) for each case using (3) with $n = 10$ and $\sigma = 10$. Use ±1.96 for Z resulting in a 95% interval. Then randomly select the next observation, y, for each of the 10,000 cases. Ask, how many times in 10,000 is the random y selection contained in the corresponding interval? In this execution we find that 94.72% or about 95% of the time the random y is contained within the corresponding interval. We can further calculate the coverage probability for each interval. That is for each (L, U) interval pair calculate F(U) – F(L) where F(x) is the CDF for the distribution being used (here normal, $\mu = 1000$ and $\sigma = 10$). In that calculation, we find that the mean coverage is 0.95002 or just about 95%. Therein lies a key point—the 100C% prediction interval has an average coverage probability of 100C%. In fact, only 68.9% of the time did the coverage exceed 95% in this example. This will be important when it comes to tolerance intervals.

Several variations are important. The first is that when σ is unknown—which is typical—we substitute student's t in for Z using the appropriate degrees of freedom (here 9). Doing this will increase the variation in the interval width but will have the same behavior. Using the means and standard deviations from the 10,000 simulated sets of $n=10$ and using the same future values, y, as previously, we find that in 95.01% of the cases that y was captured by the interval. In addition, the average coverage probability is 95.01% despite the variation increase in the standard deviation of coverage. Thus, the expected coverage is 95% and any random interval would capture the future y 95% of the time in repeating the process. Other metrics from the process can also be looked at such as ppm and other "6-sigma" metrics and ratios of percentiles.

Another important variation is the case where we want an interval for two or more future values. In that case we only have to adjust Z or t depending on if σ is known or unknown. There is an exact adjustment procedure that can be made to handle this case, but most practitioners use the much simpler Bonferroni adjustment. This is illustrated for the unknown σ case. If we want to construct a two sided $100(1-\alpha)\%$ prediction interval for the next k observations, then make an adjustment to the t value used to $t_{\alpha/(2k)}$ and $t_{1-\alpha/(2k)}$. For example, for the above example if we want a 95% prediction interval for the next 5 values then use $t_{0.0025}$ and $t_{0.9975}$ in the calculation. These values are ±3.69. The increase in t makes resulting intervals and inclusion probability larger for any single value but about 95% for simultaneous inclusion of 5 values. We can also construct one sided intervals by the appropriate adjustment to Z or t. This theory can also be used to construct prediction intervals for a future sample mean or variance and even a k out of n future sample case. Reference [1] has all of the details of these cases. Two papers by Hahn, [4] and [5], contain further detail and examples and variations of the normal distribution prediction interval case. In addition, these papers contain extensive references tracing the history of this topic.

Prediction intervals can also be constructed for attribute type data and using nonparametric methods. The common attribute cases include the binomial and Poisson distributions. There are methods for each of these that use a normal approximation in their formulation, and in those cases, we are usually told that the observed number of initial events in an initial sample size must be 5 or more. For example one rule of thumb is to require $np \geq 5$ and $n(1-p) \geq 5$ (some authors use 10 on this). For cases where the observed data fails to meet this criterion, we have to be careful because the normal approximation being used may compromise the prediction interval. There is a general way though that can handle any number of observations, including 0, that uses Monte Carlo simulation. This method depends of the relationship between the cumulative binomial distribution and the beta distribution. One way to portray this is shown below [6].

$$\sum_{x=0}^{r}\binom{r}{x}p^{x}(1-p)^{n-x} = 1-\int_{0}^{p}\frac{t^{r+1-1}(1-t)^{n-r-1}}{B(r+1,n-r)}dt \tag{4}$$

The left hand side is the binomial CDF and remarkably, this is cast in terms of the continuous beta distribution on the right side. In a binomial prediction interval, we have the current data, x observed events in a sample of size m. We want an interval (L, U) that we can expect a future binomial observation y to fall in a future sample of size n using some confidence C. What we need are potential values of the parameter p that could have generated the data we see now (x in m); then use those values with the new sample size n and randomly select an observation, y, from each case. The distribution of y is then used to construct the random interval for the future y. This is another fiducial Monte Carlo application.

Start by setting the right hand side of (4) equal to a random uniform on [0,1]; then solve that equation for p giving a plausible p that could have generated the data we see now (x in m). Repeat this 100,000 or more times saving the results (p_i) each time. Next, use each pair (n, p_i) to simulate a random future binomial observation y_i. Save the y_i values and use this distribution as the basis for constructing the prediction interval. It is very easy to do this using the freeware program R or even in MS Excel; below is a short R program (Exhibit 1), just the bare minimum, that will generate the distribution of y we need.

Exhibit 1: R code for the Binomial Prediction Interval Monte Carlo

```
#file name: BinomialPredictionInterval(R).txt
library(MASS)
k=100000    #INPUT: number of Monte Carlo runs
m=235       #INPUT: original sample size
r=9         #INPUT: original number of observations, "successes"
n=140       #INPUT: future sample size
a=r+1     #beta parameter
b=m-r     #beta parameter
u=runif(k,min=0,max=1)
p=qbeta(u,a,b)
y=rbinom(k,n,p) #random binomial predictions
write(y,file="d:/q/y.txt",ncolumns=1)    #saving y to a file in directory Q
write(p,file="d:/q/p.txt",ncolumns=1)    #saving p to a file in directory Q
#eof, end of file
```

When this code gets executed, the results, y, are saved to a text file, which can then be analyzed further in R or exported to say Minitab. An example given in [1] regarding a binomial prediction interval is instructive. In example 6.5.3 in [1] the following numbers are used: m = n = 1000, and x = 20 observed now. The authors quote an exact method due to Thatcher, A. R. (1962) and state that this is based on the Hypergeometric distribution and is used iteratively. The 95% prediction interval answer given in the text is [9, 35], and thus a future observation y from a future sample of n = 1000 can be expected to fall within this interval 95% of the time in repeated sampling. The large sample normal approximation method gives the interval [8, 32] and this is seen to be reasonably close to the exact method. The results from using the Monte Carlo simulation method discussed here show the following distribution of y following 100,000 cases—see Figure 5a.

The 2.5th and 97.5th percentile of this distribution will define the prediction interval using this method. In this case the interval is [10, 35], remarkably close to the exact method. Next suppose that x = 1 event has been observed in m = 500 and we want an upper 95% prediction interval for y the number of events in a future sample of n = 1000. Figure 5b shows this case using the R code shown above. There were 96,182 cases of $y \leq 11$ in the resulting set of y's making the 96th percentile approximately 11. Then there is at least 95% confidence if stating y = 11 as the upper prediction bound.

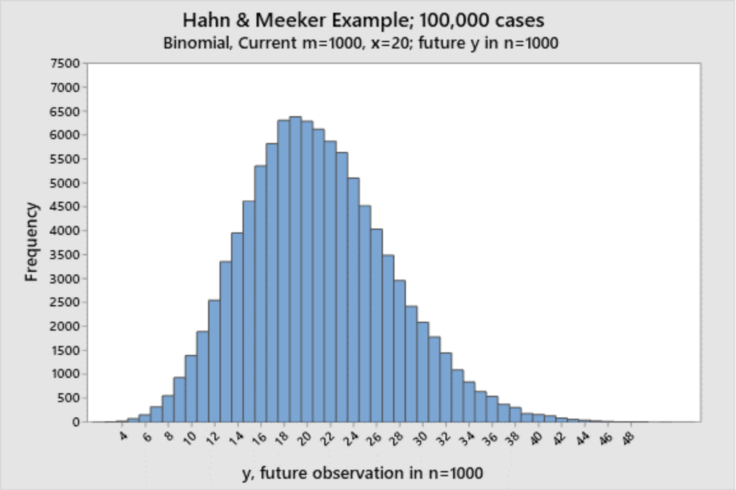


Figure 5a: Monte Carlo Binomial prediction Interval

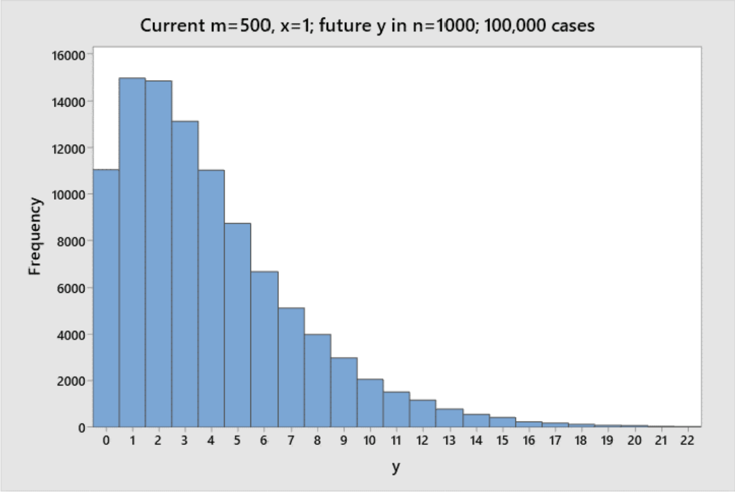


Figure 5b: Monte Carlo Binomial prediction Interval

Although the Poisson distribution has a large sample normal approximation for prediction intervals, we can use the same Monte Carlo methodology for this case by using the relationship between The Poisson CDF and the chi-square distribution [6]. That relationship is:

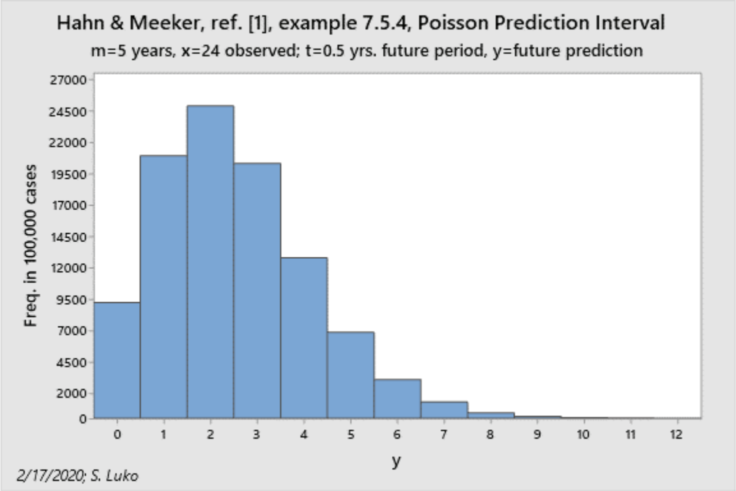
$$F(r)=\sum\nolimits_{x=0}^{r} f(x)=\int_{2\lambda s}^{\infty} g(y)\,dy \qquad (5)$$

In (5) *F(r)* is the Poisson CDF evaluated at x=r and *f(x)* is the Poisson mass function. The rate constant is λ and the observations are made on the interval of size s making the Poisson mean λs. There are x=r events observed on the interval s. The right hand integral is the upper tail chi-square distribution with 2(r+ 1) degrees of freedom. So, setting the right hand side to a random uniform variate on [0,1] one can then solve for a plausible rate λ that gave rise to the data we see (r events on s). Doing this many times generates the target distribution of λ. Then the Poisson distribution with each λ_i, together with the future interval of size t is used to select a random value of future events y. The distribution of y is then used to develop the lower and upper prediction interval confidence bounds for the future, unrealized, value of y. The following minimal R code (Exhibit 2) will generate the future values, y.

Exhibit 2: R code for the Poisson Prediction Interval Monte Carlo

```
#file name: PoissonPredictionInterval(R).txt
library(MASS)
#
k=100000          #INPUT: number of Monte Carlo observations to perform
s=5               #INPUT: original sample size or observational region size
r=24              #INPUT: original number of observed events
t=0.5             #INPUT: future sample size or observational region size
#
df=2*(r+1)        #degrees of freedom
v=rchisq(k,df)
L=v/(2*s)         #lambda
y=rpois(k,L*t)    #random observations from Poisson with mean L*t
#
write(y,file="d:/q/y.txt",ncolumns=1)   #saving y to a file in directory Q
write(L,file="d:/q/L.txt",ncolumns=1)    #saving L to a file in directory Q
#eof, end of file
```

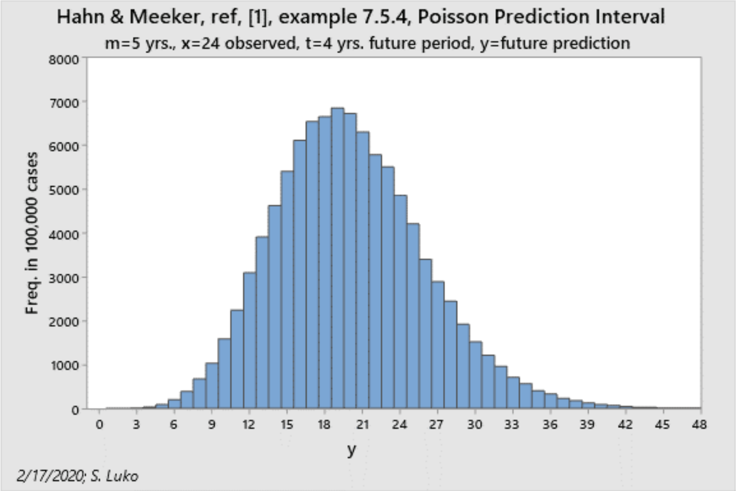
The following is another example from the Hahn and Meeker text, reference [1]. In their material on the Poisson prediction interval they give an exact formula due to Wayne Nelson (similar to the binomial case) that is used iteratively to find the upper bound for y. They also develop the large sample normal approximate methodology. In example 7.5.4 they use a case where x=24 unscheduled shutdowns have been observed over a period of s=5 years. There is a need to estimate the worst-case number of shutdowns (95% upper prediction interval) in the next 6 months. The 95% prediction upper bound value quoted is y=6 future events in a six month period. Here the next period in years is t=0.5 years. Using the Monte Carlo method, Figure 6a shows the output distribution of future values y. In these 100,000 cases, the 95th percentile is y=6 exactly matching the table values from the Hahn and Meeker tables.


Hahn & Meeker, ref. [1], example 7.5.4, Poisson Prediction Interval
m=5 years, x=24 observed; t=0.5 yrs. future period, y=future prediction

2/17/2020; S. Luko

Figure 6a: Hahn & Meeker Example 7.5.4 Poisson Prediction Interval, t=6 months

In [1] the authors continue with another variation on this example. They use the same preliminary data (x=24 on s=5 years) and construct a 4 year interval for future values, y. The quoted result for the future y is the interval [9, 34]. The Monte Carlo method generated 100,000 cases shown in figure 6b. In those cases, the proportion of values falling in the interval [9,3 4] were 0.955 or 95.5%, thus the Monte Carlo methods agrees very well with this example.


Hahn & Meeker, ref, [1], example 7.5.4, Poisson Prediction Interval
m=5 yrs., x=24 observed, t=4 yrs. future period, y=future prediction

2/17/2020; S. Luko

Figure 6b: Hahn & Meeker Example 7.5.4 Poisson Prediction Interval, t=4 years

Nonparametric (NP) cases of prediction interval estimation have numerous variations. The NP prediction interval is an interval constructed from a current sample of size m for which we want to state that in a second sample of size n at least some number r of n will be contained in the interval. We can have both one sided and 2-sided prediction intervals. We assume that the samples used are random selections from a large population or that the samples were selected from a process in a state of statistical control in the sense of Shewhart [7]. NP prediction intervals are based on the theory that any random sample of size (m) selected from any distribution, partitions that distribution into m + 1 intervals or "bins", each of which is equiprobable on average for the next observation to fall in. Another way to state this is to use the theory that for any distribution $E\{F(X_{(i)})\} = i/(n+1)$. That is, the order statistics divide a distribution into equiprobable bins on the average. For example, with m = 3 initial observations there are 4 such bins: $(-\infty, x_{(1)}]$, $[x_{(1)}, x_{(2)}]$, $[x_{(2)}, x_{(3)}]$, and $[x_{(3)}, +\infty)$. In theory, each of the four bins has an equal probability for a forth (future) observation to fall in. If a 4th value is selected, the probability that it falls within any of the 4 bins is 0.25 for each bin. The probability that it falls to the right of $x_{(1)}$, that is to say in the highest 3 bins, is 0.75. When a second value is taken, if we next want the probability that this new value should fall to the right of $x_{(1)}$ we have to consider that the sample size has changed from 3 to 4. By conditioning on the fact that the first has fallen above $x_{(1)}$ in the initial sample size of 3, we calculate the probability that the second will also fall above $x_{(1)}$. If y_1 represents the first of the additional samples and y_2 the second of the additional samples, and the initial sample size was m = 3, the formal conditional probability argument is:

$$P\{y_1 \geq x_{(1)}, y_2 \geq x_{(1)}\} = P\{y_2 \geq x_{(1)} \mid y_1 \geq x_{(1)}\} P\{y_1 \geq x_{(1)}) = (4/5)(3/4) = 0.6 \qquad (6)$$

Thus there is a 60% probability that if we select 2 new values from the same distribution as the first 3, both values will be larger than the smallest in the initial sample. All of the variations of NP prediction intervals can be analyzed using conditional probability arguments. This gets further complicated when we want to use order statistics other than the sample min or max! Here we'll focus on a few simple cases involving the sample min or max. First, consider the three commonly required intervals based on the extreme order statistics of the initial sample $X_{(1)}$ and $X_{(n)}$.

Type 1, One sided interval, case 1: $[X_{(1),}\ \infty)$

Type 1, One sided interval, case 2: $(-\infty,\ X_{(n)}]$

Type 2, Two sided interval: $[X_{(1),}\ X_{(n)}]$

If we want to use the sample extremes (min and/or max) in a sample of size n for bounding a future single observation, then the following table shows how this works.

Table 3: For an initial sample size n, the interval contains or includes a future single observation with confidence C. *Note—all subscripted variables in this table denote order statistics*

Interval	Confidence	Interval	Confidence
$(X_{(1)}, +\infty)$	$\frac{n}{n+1}$	$(-\infty, X_{(n)})$	$\frac{n}{n+1}$
$(-\infty, X_{(i)})$	$\frac{i}{n+1}$	$(X_{(i)}, +\infty)$	$\frac{n-i+1}{n+1}$
$(X_{(i)}, X_{(j)})$ $1 \leq i < j \leq n$	$\frac{j-i}{n+1}$	$(X_{(1)}, X_{(n)})$	$\frac{n-1}{n+1}$

Suppose n = 31, what is the confidence that the next observation, y, will be greater than the current sample min? The above table shows this interval to be 31/(31 + 1) or about 96.87%. If we want to claim that the sample min and max will capture the next observation in this case, the confidence is (31 – 1)/(31 + 1) or about 93.75%. Consider if we wanted to claim that the sample extremes would contain the next observation, what sample size would give 95% confidence? Use (n – 1)/(n + 1) = C. Solving for n, find that n = (1 + C)/(1 – C). If C = 0.95 is the desired confidence, then n = 39 is required. For more than one future observation to be contained as a one sided interval, it is easy to show that C = n/(n + k) is the confidence in an interval containing the next k future observations. In this case the intervals are one sided (type 1, see above). For the two sided interval (type 2, see above), we can still use a conditional probability argument similar to what was used previously. In that case when we carefully formulate this case using conditional probability, the confidence formula for the type 2 inclusion of the next k observations is:

$$C = \frac{n(n-1)}{(n+k)(n+k-1)} \tag{7}$$

As an example, suppose a sample of size n = 35 and we wish to use the sample extremes as a prediction interval for the next k = 4 observations. Equation (7) says that the confidence we can have in this interval is approximately 80%. If we must have 90% confidence, then we can use (6) with C = 0.9, and k = 4 and solve for n; or we can use trial and error. A few trials shows that n = 75 will just make C equal to 90%. In using this theory we have to be careful about outliers. Although these intervals are distribution free, legitimate outliers can and do occur and can trump the assumption of stability required for this application to work correctly.

Tolerance Intervals

Tolerance intervals are statements about the entire population or process output from a stable process. Intervals can be one or two sided, and be parametric or nonparametric. In the two sided case, an interval (a,b) means that a proportion, p, of all future values, X, from the population/process will fall within the interval with some stated confidence C. Thus, we have to specify two numbers for this type of interval, p and C. For a normal distribution, a two sided tolerance interval takes the form $\bar{x} \pm ks$ where $\bar{x}$ and s are the mean and standard deviation from a random sample of n observations, and k is the tolerance factor looked up in a table as a function of n, C and p. The case of the normal distribution is well developed and, although complex to compute, numerous resources are available for this including tables, approximate formulas, applets, standards and software applications. The software program Minitab has included tolerance intervals in the last several versions of the program. Reference [1] contains extensive tables for single and two sided case. A short table appears below.

Table 4: Table 8 extracted from ASTM E2586-19, *Standard Practice for Calculating and Using Basic Statistics*

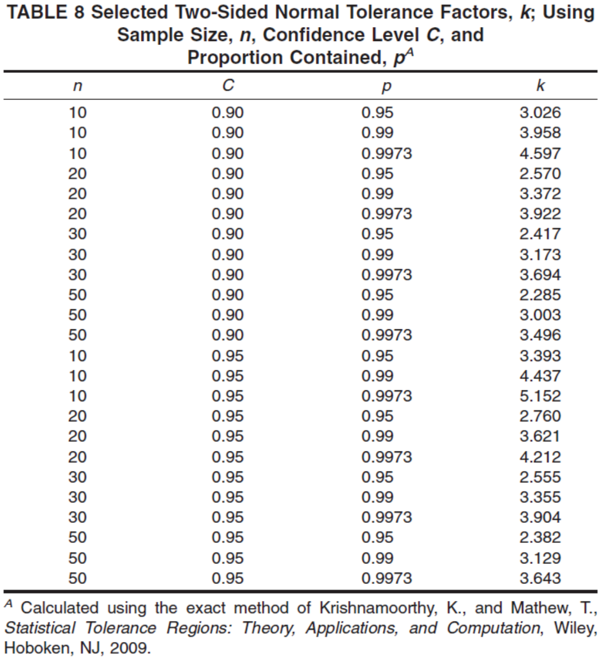


TABLE 8 Selected Two-Sided Normal Tolerance Factors, *k*; Using Sample Size, *n*, Confidence Level *C*, and Proportion Contained, *p*[A]

n	*C*	*p*	*k*
10	0.90	0.95	3.026
10	0.90	0.99	3.958
10	0.90	0.9973	4.597
20	0.90	0.95	2.570
20	0.90	0.99	3.372
20	0.90	0.9973	3.922
30	0.90	0.95	2.417
30	0.90	0.99	3.173
30	0.90	0.9973	3.694
50	0.90	0.95	2.285
50	0.90	0.99	3.003
50	0.90	0.9973	3.496
10	0.95	0.95	3.393
10	0.95	0.99	4.437
10	0.95	0.9973	5.152
20	0.95	0.95	2.760
20	0.95	0.99	3.621
20	0.95	0.9973	4.212
30	0.95	0.95	2.555
30	0.95	0.99	3.355
30	0.95	0.9973	3.904
50	0.95	0.95	2.382
50	0.95	0.99	3.129
50	0.95	0.9973	3.643

[A] Calculated using the exact method of Krishnamoorthy, K., and Mathew, T., *Statistical Tolerance Regions: Theory, Applications, and Computation*, Wiley, Hoboken, NJ, 2009.

It is interesting to note that practitioners often use the interval $\bar{x} \pm 3s$ as a quick study of process capability regardless of the sample size. Many would then go on to claim that 99.73% of the process output would be contained within the "three sigma" interval - just using the usual normal distribution 3-sgma rule and the interval as if $\bar{x}$ and s were the true μ and σ. But there is little confidence in that statement for modest sample size. In fact, for n = 30, a common sample size, the statement would be correct only about 41% of the time. Table 2 shows the factors k that should be used to make such a statement with 90% or 95% confidence. For 90%, k = 3.69 and for 95% k = 3.90 should be used.

An interesting variation for the two sided case comes about if σ is assumed known. This greatly simplifies the analysis of factors, k, in the single sided case. The derivation uses basic principles and is not all that difficult, but details are used sparingly here. A short how to summary and table of a few common cases for n, p and confidence C are shown below.

Exhibit 3: Construction of single sided Normal distribution tolerance Intervals with σ assumed known or given.

a) Select p, the minimum proportion contained in the interval, and C, the confidence coefficient for the tolerance interval. The single sided bound to be determined and is of the form $\bar{x} - k\sigma$ or $\bar{x} + k\sigma$ depending on if a lower or upper bound is the requirement.

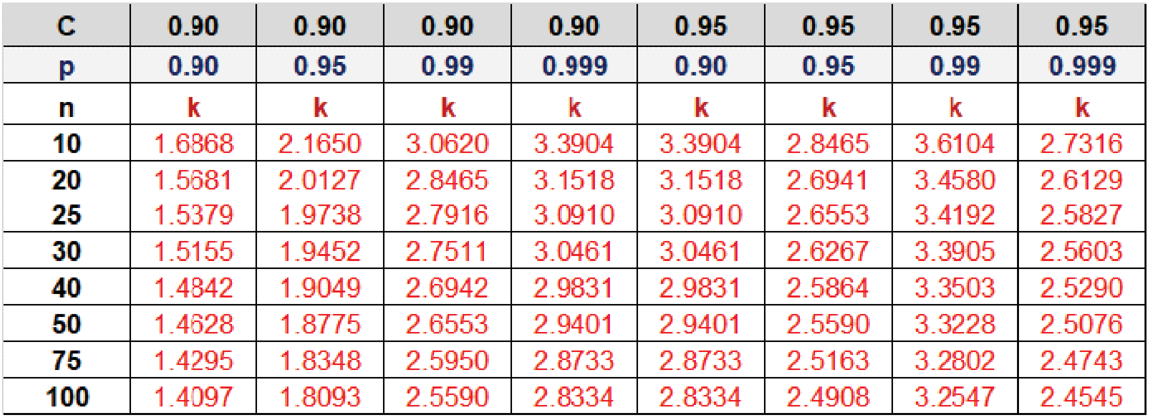
b) For the lower bound, let Z_0 be a standard normal quantile such that $P(Z > Z_0) = 1 - p$.

c) Associate with C a standard normal quantile Z_c such that $P(Z \le Z_c) = C$.

d) Then $k = Z_0 + Z_c / \sqrt{n}$

Table 5 shows selected examples of this.

Table 5: Selected Values, k, for constructing one-sided normal tolerance intervals with known standard deviation (p = proportion contained, C = confidence)



C	0.90	0.90	0.90	0.90	0.95	0.95	0.95	0.95
p	0.90	0.95	0.99	0.999	0.90	0.95	0.99	0.999
n	k	k	k	k	k	k	k	k
10	1.6868	2.1650	3.0620	3.3904	3.3904	2.8465	3.6104	2.7316
20	1.5681	2.0127	2.8465	3.1518	3.1518	2.6941	3.4580	2.6129
25	1.5379	1.9738	2.7916	3.0910	3.0910	2.6553	3.4192	2.5827
30	1.5155	1.9452	2.7511	3.0461	3.0461	2.6267	3.3905	2.5603
40	1.4842	1.9049	2.6942	2.9831	2.9831	2.5864	3.3503	2.5290
50	1.4628	1.8775	2.6553	2.9401	2.9401	2.5590	3.3228	2.5076
75	1.4295	1.8348	2.5950	2.8733	2.8733	2.5163	3.2802	2.4743
100	1.4097	1.8093	2.5590	2.8334	2.8334	2.4908	3.2547	2.4545

As an example, suppose we have a sample of n = 40 observations from a normal distribution with known/assumed σ = 12 and $\bar{x}$ = 852. If we want to construct a lower bound for 99.9% of the population/process with 90% confidence, use k = 2.9831. Then that lower bound is calculated as 842–2.9831(12) = 806.2. That interval would contain at least 99.9% of future values above the limit—with 90% confidence.

The non-parametric case for constructing tolerance intervals uses the order statistics in a sample of n observations as the bases of the tolerance interval limits. Basic formulas are available for the case of using the extremes (min and/or max) as the bounding limits for the calculation. There are several basic cases.

For the case of $[X_{(1)}, +\infty)$ or $(-\infty, X_{(n)}]$ we have essentially had a success run of size n at or above the smallest order statistic or at or below the largest order statistic. If we want to claim that at least a proportion p is greater (smaller) than or equal to $X_{(1)}$ $(X_{(n)})$ we have a success run of length n. This works like a binomial with probability p and n successes. The following is the relationship among p, C and n.

$$p^n \geq 1 - C \tag{8}$$

Equation (8) May be solved for either p, n or C. This gives:

$$p \geq \sqrt[n]{1-C}, \qquad C \geq 1 - p^n, \qquad n \geq \frac{\ln(1-C)}{\ln(p)} \tag{9a, b, c}$$

For the question of sample size, use equation (9c). Should we want to use 95% confidence and claim that a proportion of at least p = 0.99 lies above $x_{(1),}$ then using (9c) we find that n = 299 will just achieve this. We find that a number of standards used in the materials world use this sample size and for just this reason. When C = 0.9 and p = 0.9, n of 22 will work. This is a common sample size used in the automotive industry and elsewhere. Note that the two versions of the one sided case are identical in this analysis and that these are frequentest intervals—no prior Bayesian information is being used. For the case $[x_{(1)}, x_{(n)}]$, at least 100p% of the population lies in the interval with confidence C, when a sample size of n is used. Analysis of this case uses the theory of order statistics from a U(0,1) distribution. References for this include [8] and [9]. In [8] the author traces this result to the statistician S. S. Wilks [9] and further states that Shewhart put the tolerance interval idea in his head! The following equation solves the problem.

$$np^{n-1} - (n-1)p^n \geq 1 - C \tag{10}$$

In (10) we find the relationship among sample size, n, confidence, C, and proportion captured, p, when considering the interval as $[x_{(1)}, x_{(n)}]$. We can solve (10) by iteration for the unknown when any two of n, C and p are specified in advance. For example, when p = 0.99 and C = 0.9 we find that n = 388 will just make (10) true. If n = 100 is used and C = 0.95, solving we find that p = 0.9534 is the largest proportion of the population we can claim is captured by the interval $[x_{(1)}, x_{(n)}]$. It is also possible to extend tolerance intervals based on order statistics to an arbitrary interval based on any two order statistics. This is further discussed in [8] and [9].

Summary

Interval estimation is a vast area in statistical methods. In this article we have tried to expose a small sample of the many techniques available under this topic including parametric and nonparametric methods and techniques for attribute type data. It is unfortunate that the principle interval estimation method used by many practitioners is the simple confidence interval for means and proportions and that this is often used for "all occasions". So many other, more appropriate, interval methods are available! Readers are encouraged to pursue a few of the references cited, particularly [1] and [5]. For those wanting detail, [9] contains many gems.

REFERENCES:

1. Hahn and Meeker, *Statistical Intervals, A Guide for Practitioners*, Wiley Interscience, New York NY, 1991.

2 Neyman, Jerzy, *Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability,* Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences. 236 (767): 333–380, 1937.

3. Wang, Y. H., Fiducial Intervals, What are They?, The American Statistician, May 2000, Vol. 54, No. 2.

4. Hahn, Gerald J., Factors for Calculating Two-Sided Prediction Intervals for Samples From a Normal Distribution, Journal of the American Statistical Association, 64:327, 878–888, 1969

5. Hahn, Gerald J., Finding an Interval for the Next Observation from a Normal Distribution, Journal of Quality Technology, Vol. 1, No.3, July 1969

6. Evans, Hastings & Peacock, Statistical Distributions, 3rd edition, John Wiley & Sons, New York NY, 2000

7. Shewhart, Walter, Economic Control of Quality of Manufactured Product, 50th Anniversary Commemorative edition, 1980, ASQ press (originally published 1931, D. Van Nostrand Company)

8. Duncan, Acheson J., Quality Control and Industrial Statistics, Irwin, Homewood Ill., 5th edition, 1986

9. Wilks, S. S., Mathematical Statistics, Wiley & Sons Inc. 1962

MINI PAPER

Melvin Alexander, Analytician

Statistical Model Comparison Predicting Signs of Penetrating Abdominal and Pelvic Injuries using R

Abstract

This presentation takes data from a medical radiology case study of Saksobhavivat et al. (2016) designed to detect signs from Multidetector Computed Tomography (MDCT) imaging in diagnosing and treating traumatic penetrating abdominal and pelvic injuries (PAPI).

Machine Learning techniques of Random Forests, Extreme Gradient Boosting, Stepwise Logistic, and Penalized Regression using the R programming language environment were used to compare these statistical models to determine the strongest signs as indicators of diagnosing PAPI following penetrating abdominal injury.

I will demonstrate the R statistical programming language to create several machine-learning statistical models.

Results of the image analyses helped radiologists and clinicians discriminate patients requiring surgery, observation, or non-operative management.

Introduction

Penetrating abdominal and pelvic injuries (PAPI) are uncommon, potentially life-threatening, trauma injuries resulting from stab wounds, gunshot wounds (GSW), or other types.

Multidetector computed tomography (MDCT), along with advances in other medical imaging technologies has made this method one of the major modes of emergency management of PAPI for injury detection and severity, replacing the need for conducting unnecessary surgical explorations and physical examination of traumatic PAPI patients.

Determining the accuracy of MDCT, when compared to the gold-standard surgical findings, has been an important and challenging area of study. Multiple studies have looked at the overall accuracy of triple contrast CT versus single-contrast CT assessment of penetrating torso or abdominopelvic trauma, especially as it pertains to detection of PAPI.

This presentation uses the R statistical language to identify key signs that are indicators of PAPI from MDCT using modern machine learning techniques.

Materials and Methods

The University of Maryland Medical Center's Institutional Review Board (IRB) approved the prospective observational study. The written waiver of informed consent complied with the Health Insurance Portability and Accountability (HIPPA) regulations by the IRB.

Triple contrast (oral, rectal, and intravenous) MDCT has been commonly used as the primary means of evaluating penetrating abdominal and pelvic resulting from gunshot or stab wounds. Contrast media increases visibility of internal abdominal structures in CT imaging. Few studies have reported with high accuracy that triple contrast CT predicts the need for surgical treatment of penetrating abdominal and pelvic injuries.

CT images of 171 patients underwent MDCT imaging for surgery (77/171, 45.0%) or clinical follow-up (94/171, 55.0%) between October 2011—April 2013 at the University of Maryland Medical Center's (UMMC) Shock Trauma Center. The images were interpreted by three independent radiologists, (one attending radiologist and two secondary readers). Each radiologist (Column **No** in Table 1) interpreted each patient's scan and recorded findings on dedicated worksheets (Figure 2), blind to each other's imaging, clinical data, or patient's management outcomes.

Sixteen signs have been cited in the medical literature as key signs indicating PAPIs. Direct, primary signs that indicated GastroIntestinal (GI) injury included: GI wall discontinuity (Q7), subjective GI wall thickening (Q8), intramural air (Q4), bleeding into GI lumen (Q14), leakage of enteric contrast material (Q6), visible leakage of any GI content (Q5), if enteric contrast was not present at the injury site), and visible penetrating wound track (Q15) outlined by hemorrhage, air, and/or ballistic fragments) that extended up to the GI wall.

Indirect, secondary CT signs that were also evaluated included: any evidence of peritoneal violation (Q1), retroperitoneal violation (Q2), free intraperitoneal/retroperitoneal gas *adjacent to* the GI injury site (Q3a), free intraperitoneal/retroperitoneal gas *remote to* the GI injury site (Q3b), peritoneal thickening or enhancement (Q12), co-existing penetrating injuries to intraperitoneal solid organs (Q13), free intraperitoneal fluid (Q9), mesenteric hematoma (Q10), and active mesenteric hemorrhage (Q11). The 17th overall CT diagnosis of GI injury (CToverall) rated the degree of overall confidence for the presence (or absence) of a PAPI. All signs used a 5-point confidence scale (1-definitely absent, 2-may be present but unlikely, 3-unequivocal, 4-likely present, 5-definitely present).

Cross-validation (i.e., the approach of avoiding overfitting that leads to poor prediction responses) divided the full dataset (513 observations in Table 1) into training and test data tables using a randomized 80:20 split. Training data (411 or 80% of all observations, **Validation** = 0) built the regression models. Test data (102 or 20% of the remaining observations, **Validation** = 1) assessed the predictive accuracy of the regression models from the training data with confusion matrices of the correct and misclassifications.

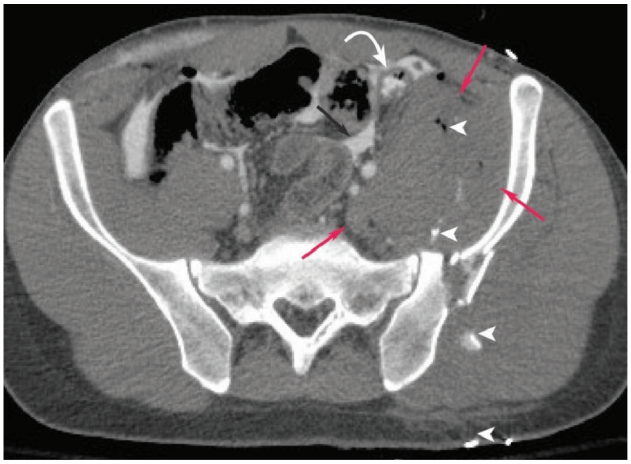


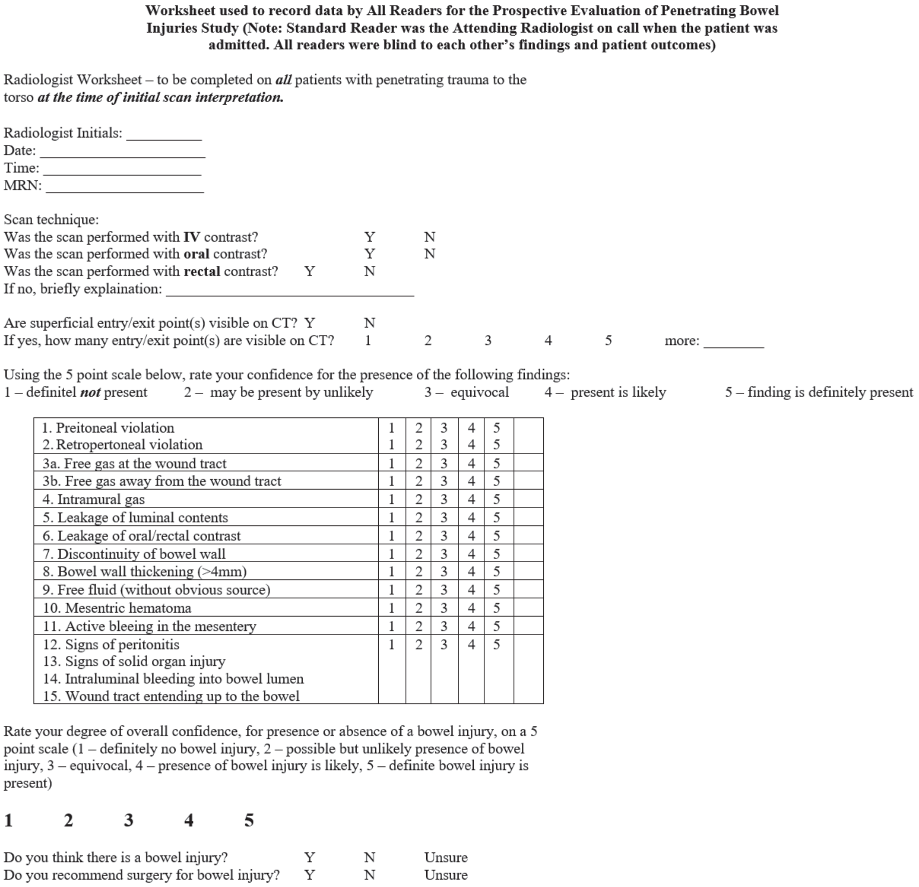
Figure 1: Image from Saksobhavivat et al. [1] of Penetrating Abdominal Pelvic Injury (PAPI) (Reprinted with permission from the European Society of Radiology)

Figure 1 of Saksobhavivat et al. [1] showed an Axial (top-down) cross-sectional view of a gunshot wound to the left pelvis with rectal contrast material extravasation. Extravasation was where contrast media leaked into surrounding tissues. Reformatted CT images demonstrated a wound tract (Q15, arrowheads) outlined by hematoma, bullet and bone fragments. There was both a descending colon (Q13, curved arrow) and jejunal wall (Q8, white arrows) thickening. Rectal contrast material (Q6, red arrows) extravasation was seen throughout the peritoneum.

Hematoma is swelling of clotted blood within tissues. The jejungal wall makes up 20 percent of the small intestines and is used to evaluate the small bowel during follow-through evaluation (a.k.a. SBFT).

The arrows and arrow heads point to certain signs that radiologists saw that help diagnose patients and determine effective medical treatments.

R has available many cutting-edge, machine learning tools like Random Forests, Extreme Gradient Boosting (XGBOOST), Penalized LASSO and Stepwise Logistic regression. See Hastie, Tibshirani, Friedman [5] for more information.



Worksheet used to record data by All Readers for the Prospective Evaluation of Penetrating Bowel Injuries Study (Note: Standard Reader was the Attending Radiologist on call when the patient was admitted. All readers were blind to each other's findings and patient outcomes)

Radiologist Worksheet – to be completed on ***all*** patients with penetrating trauma to the torso ***at the time of initial scan interpretation.***

Radiologist Initials: __________
Date: ____________________
Time: ____________________
MRN: ____________________

Scan technique:
Was the scan performed with **IV** contrast? Y N
Was the scan performed with **oral** contrast? Y N
Was the scan performed with **rectal** contrast? Y N
If no, briefly explaination: ______________________________

Are superficial entry/exit point(s) visible on CT? Y N
If yes, how many entry/exit point(s) are visible on CT? 1 2 3 4 5 more: ________

Using the 5 point scale below, rate your confidence for the presence of the following findings:
1 – definitel ***not*** present 2 – may be present by unlikely 3 – equivocal 4 – present is likely 5 – finding is definitely present

1. Preitoneal violation	1	2	3	4	5	
2. Retropertoneal violation	1	2	3	4	5	
3a. Free gas at the wound tract	1	2	3	4	5	
3b. Free gas away from the wound tract	1	2	3	4	5	
4. Intramural gas	1	2	3	4	5	
5. Leakage of luminal contents	1	2	3	4	5	
6. Leakage of oral/rectal contrast	1	2	3	4	5	
7. Discontinuity of bowel wall	1	2	3	4	5	
8. Bowel wall thickening (>4mm)	1	2	3	4	5	
9. Free fluid (without obvious source)	1	2	3	4	5	
10. Mesentric hematoma	1	2	3	4	5	
11. Active bleeing in the mesentery	1	2	3	4	5	
12. Signs of peritonitis	1	2	3	4	5	
13. Signs of solid organ injury						
14. Intraluminal bleeding into bowel lumen						
15. Wound tract entending up to the bowel						

Rate your degree of overall confidence, for presence or absence of a bowel injury, on a 5 point scale (1 – definitely no bowel injury, 2 – possible but unlikely presence of bowel injury, 3 – equivocal, 4 – presence of bowel injury is likely, 5 – definite bowel injury is present)

1 2 3 4 5

Do you think there is a bowel injury? Y N Unsure
Do you recommend surgery for bowel injury? Y N Unsure

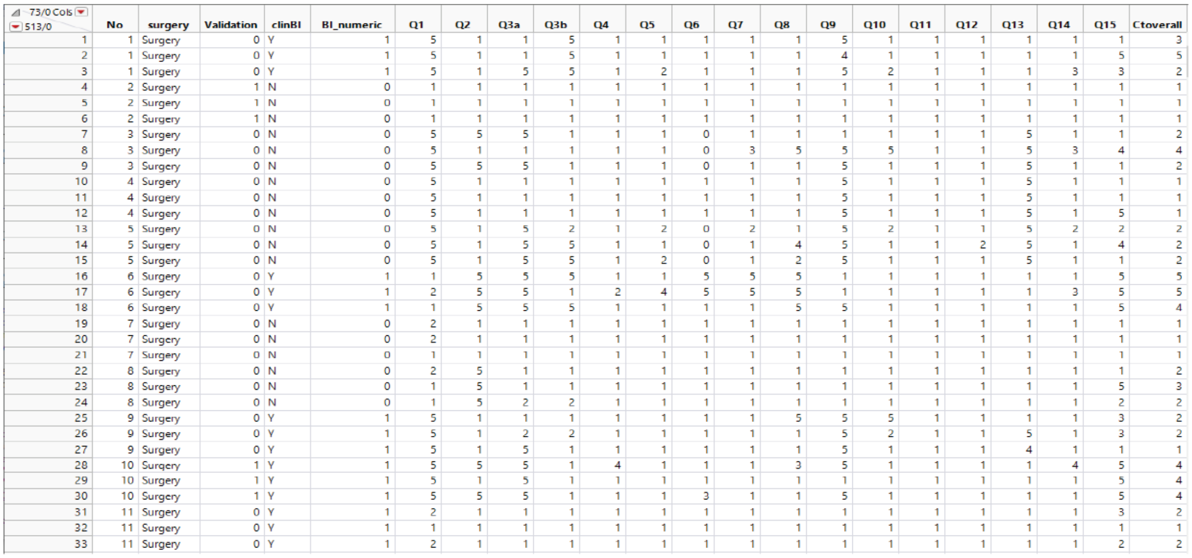
Figure 2: Worksheet with the Ordinal Scale used by independent radiologists to record data (Reprinted with permission from the Author)

The ordinal scale was adopted because:

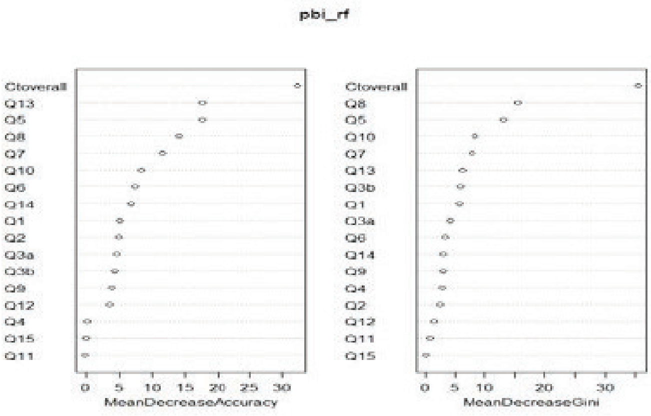
- past reviews of the literature only related signs to binary response outcomes (1 = PAPI presence, 0 = PAPI absence)
- it provided an improved measurement scale that extended beyond nominal, two-level outcomes of previous studies

- the scale was like other scales I helped develop in other studies
- careful training was given to readers so that they could use the rating scale to score CT scans consistently and correctly (using reference images of "known" signs that appeared on images)
- this scale was useful for assessing inter-reader reliability, reproducibility or variability among of image readers

Table 1: Selected Records of the 513 ReaderData Dataset for Analysis



73/0 Cols 513/0	No	surgery	Validation	clinBI	BI_numeric	Q1	Q2	Q3a	Q3b	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Ctoverall
1	1	Surgery	0	Y	1	5	1	1	5	1	1	1	1	1	5	1	1	1	1	1	1	3
2	1	Surgery	0	Y	1	5	1	1	5	1	1	1	1	1	4	1	1	1	1	1	5	5
3	1	Surgery	0	Y	1	5	1	5	5	1	2	1	1	1	5	2	1	1	1	3	3	2
4	2	Surgery	1	N	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	2	Surgery	1	N	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	2	Surgery	1	N	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	3	Surgery	0	N	0	5	5	5	1	1	1	0	1	1	1	1	1	1	5	1	1	2
8	3	Surgery	0	N	0	5	1	1	1	1	1	0	3	5	5	5	1	1	5	3	4	4
9	3	Surgery	0	N	0	5	5	5	1	1	1	0	1	1	5	1	1	1	5	1	1	2
10	4	Surgery	0	N	0	5	1	1	1	1	1	1	1	1	5	1	1	1	5	1	1	1
11	4	Surgery	0	N	0	5	1	1	1	1	1	1	1	1	5	1	1	1	5	1	1	1
12	4	Surgery	0	N	0	5	1	1	1	1	1	1	1	1	5	1	1	1	5	1	5	1
13	5	Surgery	0	N	0	5	1	5	2	1	2	0	2	1	5	2	1	1	5	2	2	2
14	5	Surgery	0	N	0	5	1	5	5	1	1	0	1	4	5	1	1	2	5	1	4	2
15	5	Surgery	0	N	0	5	1	5	5	1	2	0	1	2	5	1	1	1	5	1	1	2
16	6	Surgery	0	Y	1	1	5	5	5	1	1	5	5	5	1	1	1	1	1	1	5	5
17	6	Surgery	0	Y	1	2	5	5	1	2	4	5	5	5	1	1	1	1	1	3	5	5
18	6	Surgery	0	Y	1	1	5	5	5	1	1	1	1	5	5	1	1	1	1	1	5	4
19	7	Surgery	0	N	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	7	Surgery	0	N	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
21	7	Surgery	0	N	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
22	8	Surgery	0	N	0	2	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
23	8	Surgery	0	N	0	1	5	1	1	1	1	1	1	1	1	1	1	1	1	1	5	3
24	8	Surgery	0	N	0	1	5	2	2	1	1	1	1	1	1	1	1	1	1	1	2	2
25	9	Surgery	0	Y	1	5	1	1	1	1	1	1	1	5	5	5	1	1	1	1	3	2
26	9	Surgery	0	Y	1	5	1	2	2	1	1	1	1	1	5	2	1	1	5	1	3	2
27	9	Surgery	0	Y	1	5	1	5	1	1	1	1	1	1	5	1	1	1	4	1	1	1
28	10	Surgery	1	Y	1	5	5	5	1	4	1	1	1	3	5	1	1	1	1	4	5	4
29	10	Surgery	1	Y	1	5	1	5	1	1	1	1	1	1	1	1	1	1	1	1	5	4
30	10	Surgery	1	Y	1	5	5	5	1	1	1	3	1	1	5	1	1	1	1	1	5	4
31	11	Surgery	0	Y	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	2
32	11	Surgery	0	Y	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
33	11	Surgery	0	Y	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2



```
          MeanDecreaseGini Variables
Ctoverall       35.4559999 Ctoverall
Q8              15.5728002        Q8
Q5              13.1841609        Q5
Q10              8.3487603       Q10
Q7               7.7890653        Q7
Q13              6.1821116       Q13
Q3b              5.7022708       Q3b
Q1               5.5847796        Q1
Q3a              4.0638853       Q3a
Q6               3.3230109        Q6
Q14              2.9069360       Q14
Q9               2.8594824        Q9
Q4               2.6959956        Q4
Q2               2.4053300        Q2
Q12              1.3662419       Q12
Q11              0.7859101       Q11
Q15              0.0000000       Q15
```

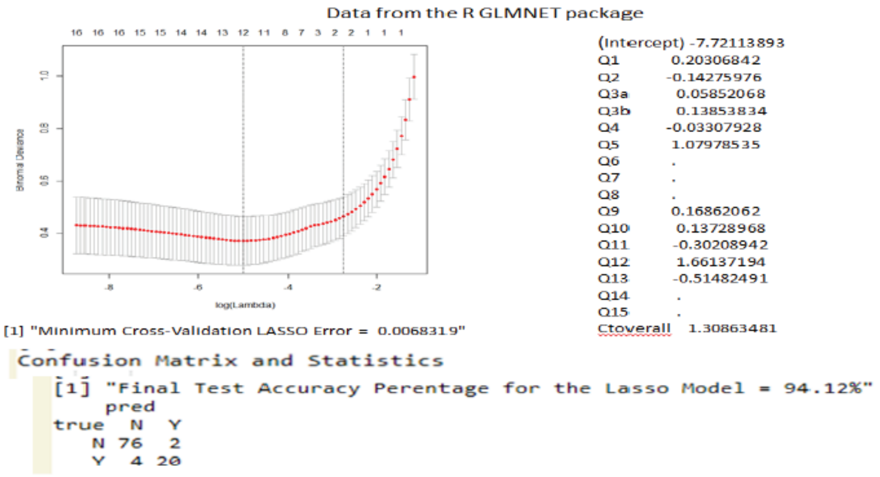
```
Confusion Matrix and Statistics

          Reference
Prediction  N  Y
         N 77  6
         Y  1 18
```

```
               Accuracy : 0.9314
                 95% CI : (0.8637, 0.972)
    No Information Rate : 0.7647
    P-Value [Acc > NIR] : 8.156e-06

                  Kappa : 0.7945

 Mcnemar's Test P-Value : 0.1306

            Sensitivity : 0.7500
            Specificity : 0.9872
         Pos Pred Value : 0.9474
         Neg Pred Value : 0.9277
             Prevalence : 0.2353
         Detection Rate : 0.1765
   Detection Prevalence : 0.1863
      Balanced Accuracy : 0.8686

       'Positive' Class : Y
```

Figure 3: Importance Plot of the Random Forest

Figure 3 shows the Variable importance plot for pbi_rf (PAPI response variable) identifying the top six variables (Ctoverall, Q5, Q7, Q8, Q10, and Q13) based on Model Accuracy (MeanDecreaseAccuracy) and Gini (MeanDecreaseGini) value. The table on the right listed variables in decreasing order of importance based on a measure (MeanDecreaseGini for node or sign impurity).

The Confusion Matrix showed the predictive performance (Accuracy of 0.9314 or 93.14%) of the Random Forest Model on the Test dataset.



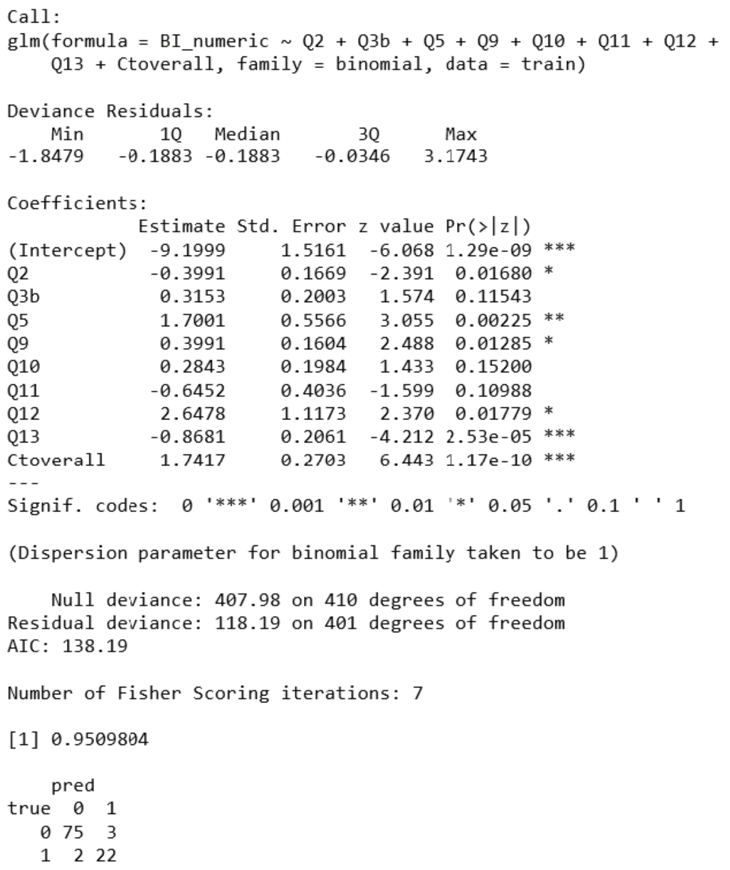
Data from the R GLMNET package

```
(Intercept) -7.72113893
Q1           0.20306842
Q2          -0.14275976
Q3a          0.05852068
Q3b          0.13853834
Q4          -0.03307928
Q5           1.07978535
Q6           .
Q7           .
Q8           .
Q9           0.16862062
Q10          0.13728968
Q11         -0.30208942
Q12          1.66137194
Q13         -0.51482491
Q14          .
Q15          .
Ctoverall    1.30863481
```

```
[1] "Minimum Cross-Validation LASSO Error =  0.0068319"
Confusion Matrix and Statistics
   [1] "Final Test Accuracy Perentage for the Lasso Model = 94.12%"
       pred
   true  N  Y
      N 76  2
      Y  4 20
```

Figure 4: Cross-validation Error (Binomial Deviance) by Log(lambda) Plot and Regression Coefficients from the LASSO Model, assessing the model accuracy against the Validation Test Data from the R GLMNET package

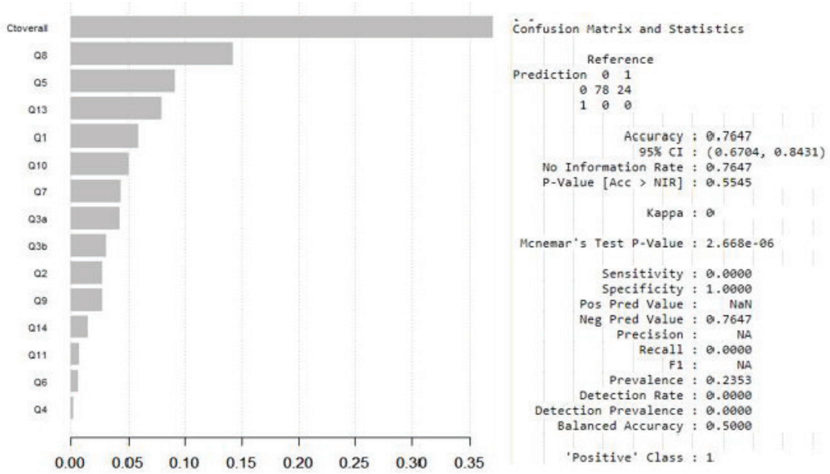
Figure 4 plots the cross-validation error according to the log of lambda. The left dashed vertical line indicates that the log of the optimal value of lambda is approximately –5, which is the one that minimizes the prediction error. This lambda value gave the most accurate model percentage of 94.12% (100* [76+20]/[76+20+4+2]=100*0.94118=94.118% ≈ 94.12%).



```
Call:
glm(formula = BI_numeric ~ Q2 + Q3b + Q5 + Q9 + Q10 + Q11 + Q12 +
    Q13 + Ctoverall, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8479  -0.1883  -0.1883  -0.0346   3.1743

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.1999     1.5161  -6.068 1.29e-09 ***
Q2           -0.3991     0.1669  -2.391  0.01680 *
Q3b           0.3153     0.2003   1.574  0.11543
Q5            1.7001     0.5566   3.055  0.00225 **
Q9            0.3991     0.1604   2.488  0.01285 *
Q10           0.2843     0.1984   1.433  0.15200
Q11          -0.6452     0.4036  -1.599  0.10988
Q12           2.6478     1.1173   2.370  0.01779 *
Q13          -0.8681     0.2061  -4.212 2.53e-05 ***
Ctoverall     1.7417     0.2703   6.443 1.17e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 407.98 on 410 degrees of freedom
Residual deviance: 118.19 on 401 degrees of freedom
AIC: 138.19

Number of Fisher Scoring iterations: 7

[1] 0.9509804

    pred
true  0  1
   0 75  3
   1  2 22
```

Figure 5: Stepwise Logistic model on the Test data resulting from R's StepAIC() option of the glm function in the MASS library.

Figure 5 shows that Ctoveral, Q2, Q3b, Q5, Q9, Q10, Q11, Q12, and Q13 were the strongest signs with a prediction accuracy of 95.02% (100 * [22+75]/ [75+22+3+2]=100* 0.9509804).



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 78 24
         1  0  0

               Accuracy : 0.7647
                 95% CI : (0.6704, 0.8431)
    No Information Rate : 0.7647
    P-Value [Acc > NIR] : 0.5545

                  Kappa : 0

 Mcnemar's Test P-Value : 2.668e-06

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.7647
              Precision :     NA
                 Recall : 0.0000
                     F1 :     NA
             Prevalence : 0.2353
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 1
```

Figure 6: Relative Importance Variables and Confusion Matrix from XGBOOST

Figure 6 shows that Ctoverall, Q1, Q8, Q5, Q13, Q1, and Q10 were the strongest signs with a prediction accuracy of 76.47% (100 * [78]/ [78+24+0+0]=100* 0.7647).

Table 2. Model Comparisons on the Test Data

Model	Confusion Matrix	Accuracy	Top Predictor signs
LASSO	Reference Prediction N Y N 76 2 Y 4 20	[1] "Final Test Accuracy Percentage for the Lasso Model = 94.12%"	Q1, Q2, Q3a, Q3b, Q4, Q5, Q9, Q10, Q11, Q12, Q13, Ctoverall
Random Forest	Reference Prediction N Y N 77 6 Y 1 18	[1] 0.9313725 or 93.14%	Ctoverall, Q8, Q5, Q10, Q7, Q13
XGBoost	Reference Prediction 0 1 0 78 24 1 0 0	Accuracy: 0.7647	Ctoverall, Q1, Q8, Q5, Q13, and Q10
Stepwise Logistic	Reference Prediction N Y N 75 3 Y 2 22	[1] 0.9509804 or 95.10%	Ctoverall, Q2,Q3b,Q5, Q9,Q10,Q11, Q12,Q13

Conclusion:

The Stepwise Logistic model performed the best on the Test Data with a prediction accuracy of 95.10%. The next best models were the penalized LASSO, and Random Forest with prediction accuracies of 94.12% and 93.14%, respectively. The worst performing model in this example is the XGBOOST.

The Stepwise Logistic model had a simpler mathematical form, was easier to interpret than the other models, and because of parsimony, was chosen as the most practical model to deploy. The strongest predictors from the Stepwise Logistic Model were: Ctoverall, Q2, Q3b, Q5, Q9, Q10, Q11, Q12, and Q13. Signs Q4, Q2, and Q10 are other indicators that radiologists and clinicians should especially notice also. The most accurate CT signs across all models were Ctoverall, Q5, Q10, and Q13. Although Saksobhavivat et al. [1] found that the combination of Q8-GI wall thickening, Q6-leakage of GI content, and Q15-wound tract extending up to the bowel wall increased diagnostic accuracy. The common signs of Ctoverall, Q5, Q10, and Q13 from all four models provide additional indicators that radiologists should look for when viewing CT scans for PAPI.

References:

1. Saksobhavivat N, Shanmuganathan K, Boscak A, et al. (2016). Diagnostic accuracy of triple-contrast multi-detector computed tomography for detection of penetrating gastrointestinal injury: a prospective study. *European Radiology*. 2016;26(11): 4107–4120.

2. Ghumman, Z., Monteiro, S, et al. (2020). Accuracy of Preoperative MDCT in Patients With Penetrating Abdominal and Pelvic Trauma, *Canadian Association of Radiologists Journal*, 2020, https://journals.sagepub.com/doi/full/10.1177/0846537119888375.

3. Alexander, M., (2016) Exploring JMP®'s Image Visualization Tools in Medical Diagnostic Applications, *Proceedings of the 2016 SouthEast SAS® Users Group (SESUG) Conference*, Bethesda, MD, October 16–18, 2016.

4. Alexander, M., (2019). Random Forest Example of the Boston Housing Data using the Base SAS® and the PROC_R macro in SAS® Enterprise Guide, Oct 23, *Proceedings of the 2019 SouthEast SAS® Users Group (SESUG) Conference*, Williamsburg, VA, October 20–22, 2019.

5. Hastie, T., Tibshirani, R., and Friedman, J., (2017) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, 12th Printing*, Springer-Verlag: New York, NY.

Upcoming Conference Calendar

1) Data Science, Statistics & Visualization 2020 Virtual Conference

Week beginning 29 July 2020

Data Science, Statistics & Visualization (2020) is a virtual conference aimed at bringing together researchers and practitioners interested in the interplay of statistics, computer science, and visualization, and to build bridges between these fields. The conference highlights contributions to practical applications, and in particular those which are linking and integrating these subject areas. Presentations will be oriented towards a very wide scientific audience and will cover topics such as machine learning, the visualization of data, big data infrastructures and analytics, interactive learning, advanced computing, and other important themes.

More information available at: https://www.samsi.info/

2) RSS International Virtual Conference

Week beginning 7 September 2020

The RSS International Conference regularly attracts more than 600 attendees from over 40 countries, providing one of the best opportunities for anyone interested in statistics and data science to come together to share knowledge and network. This year's conference will once again feature top keynote speakers and invited talk sessions but will be slightly different. Due to the impact of the Covid-19 pandemic, the RSS 2020 Conference is moving online.

More information available at: https://rss.org.uk/

3) JSM 2020 Virtual Conference

Week beginning 2 August 2020

Joint Statistical Meeting offers a unique opportunity for statisticians in academia, industry, and government to exchange ideas and explore opportunities for collaboration. Beginning statisticians (including students) can learn from and interact with senior members of the profession.

More information available at: https://ww2.amstat.org/

2020

ELECTED POSITIONS

CHAIR-ELECT/AUDITING
Matthew Barsalou
chairelect@asqstatdiv.org

CHAIR
Amy Ruiz
chair@asqstatdiv.org

PAST CHAIR/NOMINATING
Mindy Hotchkiss
pastchair@asqstatdiv.org

SECRETARY
Gary Gehring
secretary@asqstatdiv.org

TREASURER
Shoshana Bokelman
treasurer@asqstatdiv.org

APPOINTED POSITIONS

Vice Chair Content

Gary Gehring
content@asqstatdiv.org

STATISTICS DIGEST EDITOR
Harish Jose
harishjose@gmail.com

NEWSLETTER CONTENT REVIEW
Kurtis Shuler
kurtis.shuler@gmail.com

WEBINAR / YOU TUBE
Harry Rowe
webinars@asqstatdiv.org

SOCIAL MEDIA MANAGER
Brian Sersion
bsersion@gmail.com

WEBMASTER
Geoff Farmer
gfarmer118@yahoo.com

Vice Chair Community Involvement

Jennifer Williams
community@asqstatdiv.org

FTC PROGRAM REP
Richard McGrath
rnmcgra@bgsu.edu

FTC STEERING COMMITTEE
Bill Myers
myers.wr@pg.com

FTC SHORT COURSE CHAIR
Brian Sersion
bsersion@gmail.com

WCQI COORDINATOR
Steven Schuelka
pastchair@asqstatdiv.org

DATA SCIENCE/ANALYTICS INTEREST GROUP CHAIR
Michael Mladjenovic
datascience@asqstatdiv.org

MEMBERSHIP/ OUTREACH CHAIR
Jennifer Williams
membership@asqstatdiv.org

EZINE
Shoshana Bokelman
secretary@asqstatdiv.org

VOC CHAIR
Michael Kirchner
michael.r.kirchner@gmail.com

Vice Chair Awards

Peter Parker
awards@asqstatdiv.org

FTC STUDENT/EARLY CAREER GRANTS
Jennifer Williams
ftcgrants@outlook.com

NELSON AWARD
TBD

HUNTER AWARD
Joel Smith
joelmcquarriesmith@gmail.com

YOUDEN ADDRESS
Steven Schuelka
pastchair@asqstatdiv.org

BISGAARD AWARD
TBD

EXAMINING CHAIR
Daksha Chokshi
examining@asqstatdiv.org

ASQ
The Global Voice of Quality®

The *ASQ Statistics Division Newsletter* is published three times a year by the Statistics Division of the American Society for Quality.

All communications regarding this publication, **EXCLUDING CHANGE OF ADDRESS**, should be addressed to:

Harish Jose
Editor
email: newsletter@asqstatdiv.org

Other communications relating to the ASQ Statistics Division should be addressed to:

Amy Ruiz
Division Chair
email: chair@asqstatdiv.org

Communications regarding **change of address** should be sent to ASQ at:

ASQ
P.O. Box 3005
Milwaukee, WI 53201-3005

This will change the address for all publications you receive from ASQ. You can also handle this by phone (414) 272-8575 or (800) 248-1946.

Upcoming Deadlines for Submissions

Issue	Vol	No.	Due Date
October	39	3	15 August 20

VISIT THE STATISTICS DIVISION WEBSITE

www.asq.org/statistics
https://my.asq.org/communities/home/177

ASQ Periodicals with Applied Statistics content

Journal of Quality Technology
http://www.asq.org/pub/jqt/

Quality Engineering
http://www.asq.org/pub/qe/

Six Sigma Forum
http://www.asq.org/pub/sixsigma/

STATISTICS DIVISION RESOURCES

LinkedIn Statistics Division Group
https://www.linkedin.com/groups/ASQ-Statistics-Division-2115190

Scan this to visit our LinkedIn group!

Connect now by scanning this QR code with a smartphone (requires free QR app)

Check out our YouTube channel at
youtube.com/asqstatsdivision

> "A process can be routinely operated so as to produce not only product but also *information* on how to improve the process and the product."
>
> —George. E. P. Box

Statistics Division
ASQ
Excellence Through Quality™