A reflection on the historical development of the subject area of Statistics

Moudud Alam

School of Information and Engineering

Dalarna University

maa@du.se

Outline

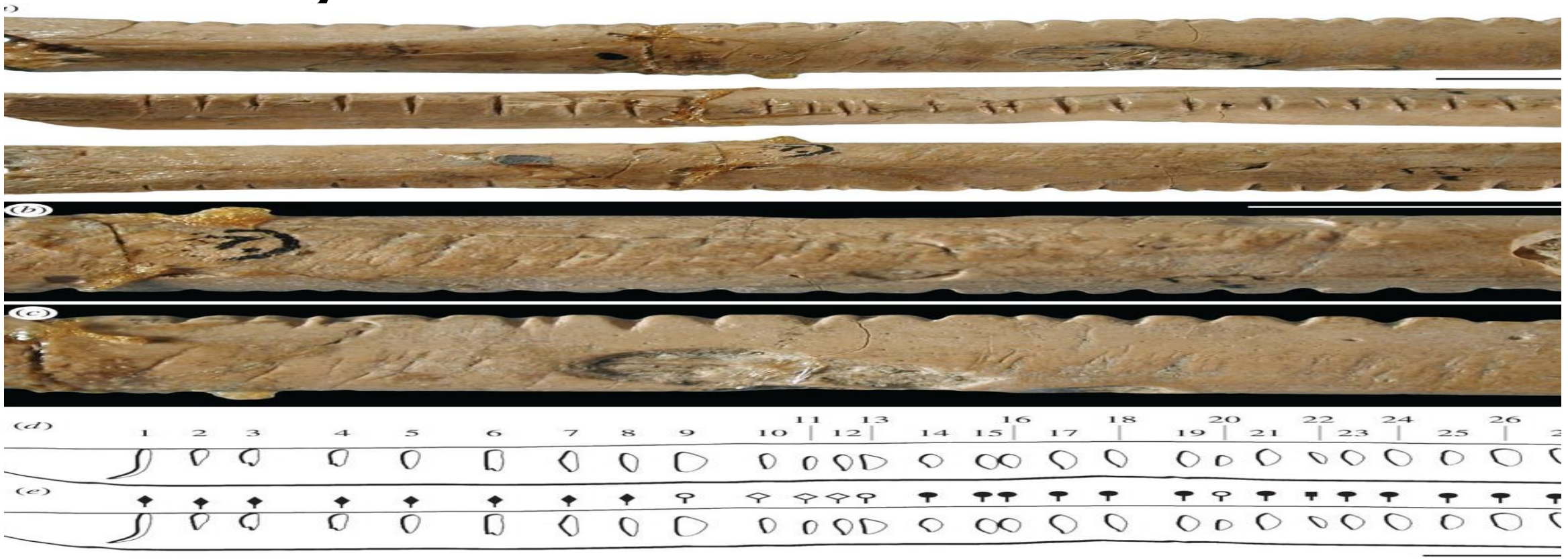
Part 1: History of Statistics and Data Science

1. Historical development of Statistics
2. Statistics and Data Science
3. Summarizing the changes

Part 2: Reflection on the Current State of Research

1. Example 1: Perils of ad hoc methods for data analysis.
2. Example 2: Own research about estimating service lifetime of traffic signs.

A 42 K yrs. old data base



Note: Baboon fibula with number marking from Border Cave located in KwaZulu-Natal, 2 km north of the Ngwavuma River and 82 km west of the Indian Ocean.

Source: d'Errico et al., 2018

Long history short

- Humans started counting (animals, people, etc.) sometimes between 40,000 and 77,000 years ago (d'Errico, et al., 2018).
- There are clear mentions of "census" at several places in the Old Testament (e.g. Number 1: 2-3; 13th century BC).
- There are traces of statistical practices in other ancient documents (e.g. Mahabharata, apprx. 8th - 3rd century BC; in works by Chanakya, 4th - 3rd century BC; Talmud, 500 AD; and in works by Al-Kindi, 9th century AD).

History of Statistics as an academic subject

- Universities around Europe started offering courses involving some statistics from mid 17th century; Hermann Conring's 1660 lecture is considered as the first one.
- Gotfrid Achenwall (1719-1772) is credited for the first use of the term "Statistik" (?) in a university lecture.
- By 1880s many universities in US started offering courses in Statistics.
- In Sweden, Gustaf Sundbärg was appointed as the first Docent (Associate professor) of Statistics at Stockholms Högskola in 1903 (?).
- Karl Pearson established the Department of Applied Statistics in 1911 at University College of London.
- Mathematical Statistics became a popular theme in late 1930's.

Development of Research in Statistics over a century

Editorial, 1838, JRSS: 1(1)	Editorial, 1901, Biometrika, 1(1)	King, I. W. (1930), Ann. Math. Stat., 1(1)
Statistics, therefore, may be said, …. to be ascertaining and bringing together of those facts which are calculated to illustrate the condition and prospects of the society.	Biometrika shall serve as a means not only of collecting under one title biological data … but also of spreading a knowledge of such statistical theory as may be requisite for their scientific treatment. … *The recent development of statistical theory, dealing with statistical data on the lines suggested by Mr. Francis Galton, has rendered it possible to deal with statistical data of every various kinds in a simple and intelligible way…*	At the time when our association (ASA) was founded, statistical methods were extremely simple. … For some time past, … membership … tending to divide into two groups – those familiar with advanced mathematics and those who have not familiarized … This journal will deal, not only with mathematical technique of statistics but also applications …

History of many separation and reunion

- From the beginning of 20th century, statistics virtually left its old descriptive nature and engaged in methods development
- Until 1920's statisticians were scattered al over in the university departments.
- During 1930's all universities started gathering statisticians under the umbrella of mathematical sciences.
- By 1960's mathematics consume the spirit of statistics when some prominent statisticians like John Tukey revolted!
- By 1980's computer intensive methods started challenging mathematical approach.

Development of the core contents of statistics education

Yule (1911), Into. Theory of Statis.	Cramer (1946), Math. Methos of Statis.	Wasserman (2004), All of Statistics
History, Notations, Descriptive statistics, Correlation and association, Basic probability distributions	Set and points, Integral/Measure theory, Random variable, Probability distribution, Inference (sampling distribution, estimation, and test)	Probability, Random variable, Parametric inference, Hypothesis testing, Bayesian inf., Decision theory, linear and non-linear models, Causal inference, Directed graph, Stochastic process, Simulation, Bootstrap, …

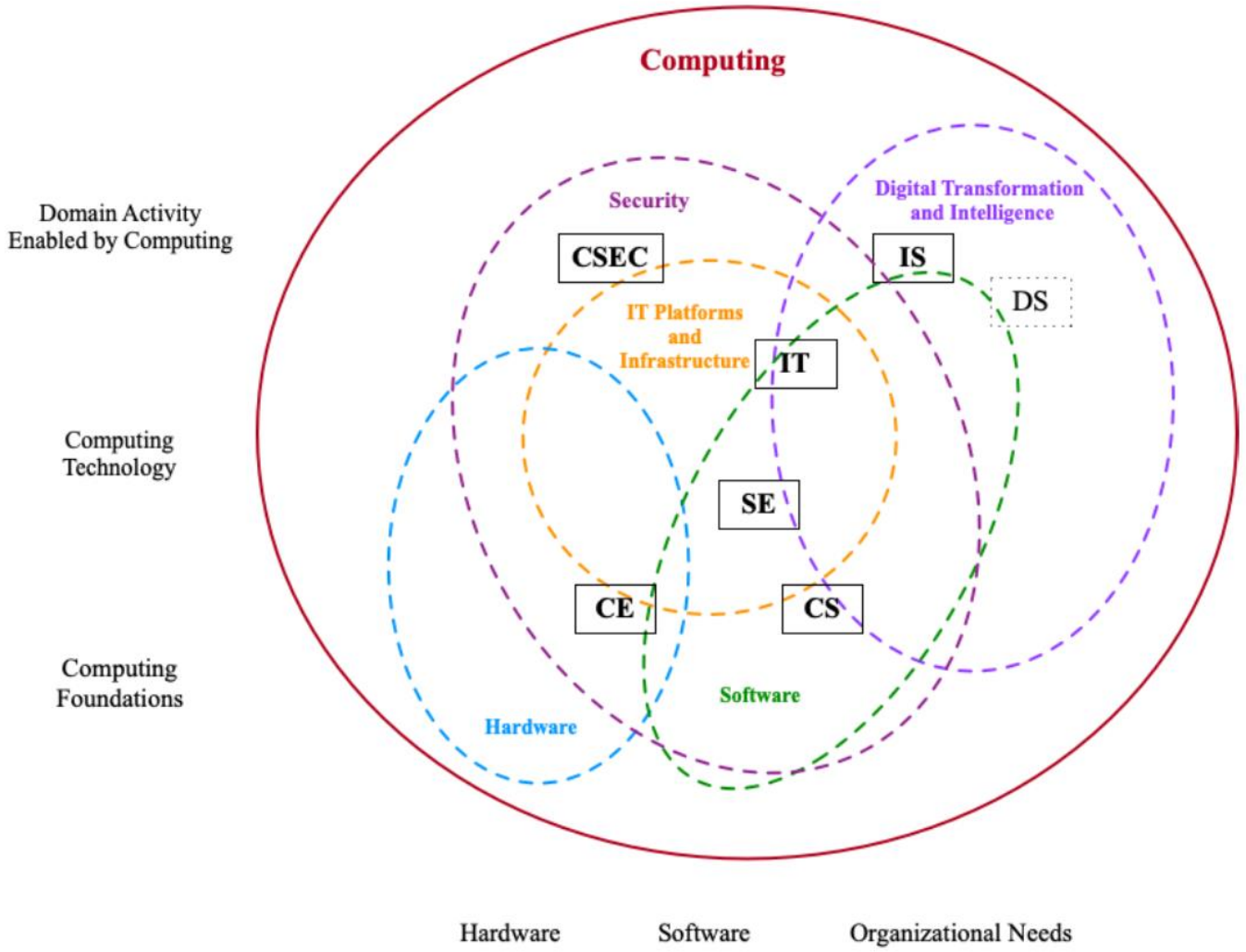
Emergence of Data Science from Statistics

- In 1965, Swedish Ministry of Education decided to open a new subject named Information Processing Specialising in Methods for Administrative Data Processing.
- Jeff Wu gave a talk “Statistics = Data Science” in 1997.
- Cleveland (2001), Data Science: An action plan for expanding… was published
- Dalarna University started “Master’s in Business Intelligence”, under the Department of Microdata Analysis, from 2011, and Data Science from 2017.
- Royal Statistical Society hosted a debate in 2015, “Data Science and Statistics: Different Worlds”?
- Data Science master’s programme started appearing in US around 2015.
- Donoho (2017), “50 Years of Data Science” was published.

Academic community initiatives for Data Science

- EDISON Project was initiated by EU in 2015.
- ASA and ACM collaboration “The Interface” changed name to “Symposium on Data Science and Statistics (SDSS)” from 2018.
- ACM Published CC2020, recognizing Data Science as an emerging subject.
- Cuadrado-Gallego and Demchenko (2020), “The Data Science Framework A View from the EDISON Project”, was published.
- ACM (2021), Computing Competencies for Undergraduate Data Science Curricula (CCDS2021), was published.
- ASA journal Teaching in Statistics became Teaching in Statistics and Data Science from 2022.

Data Science from Computing perspective (CC2020, ACM)



What changes are made in the curriculum?

Suggestion from Donoho (2017)	Dalarna University, Data Science
Data Gathering, Preparation, and Exploration	Data Collection and Data Quality, Business Intelligence,
Data Representation and Transformation	Data Warehousing
Computing with Data	Python- and R-programming
Data Modeling	Statistical Learning, Machine Learning, Spatial Data and GIS
Data Visualization and Presentation	Data Analysis and Visualization, Data Driven Leadership
Science about Data Science	
	Risk Analysis, …

What is happening in Sweden?

- Almost all the universities in Sweden has at least one programme with a term “Data science”, or alike, in its name.
- Some of this programmed are just renamed their old programmes without any major change in it (see e.g. Data Science/Business Intelligence programmes run by the Statistics Department at Örebro University) .
- Uppsala University is the most extreme:
  - Master's Programme in Data Science - Data Engineering, Department of Information Technology.
  - Master's Programme in Data Science - Image Analysis and Machine Learning, Department of Information Technology
  - Master's Programme in Data Science - Machine Learning and Statistics, Department of Information Technology
  - Master's Programme in Statistics and Data Science, Department of Statistics

What changes have occurred in the research?

- Access to extremely powerful computational facilities, e.g. high-speed computing along with capacity to handle huge data volume (from 1.44 mb FDD, to terabytes of cloud memory).
- Unconventional data sources (unstructured data, unconventional data collection method, high dimensional data or p>>>n).
- Emphasis on cross- and multi-disciplinary research.
- Wide acceptance of empirical evidence in the science of data analysis.

Perils of cross validation for variable selection

- When dealing with complicated models, we often rely on cross-validation. But how good is that?
- To illustrate the problem, I take an example from Leng, Lin, and Wahba (2006).
- Assume we have a simple linear model $y_i = \boldsymbol{X}_i\beta + \varepsilon_i$, with all usual but $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_d$, $\beta = \left(\beta_1, \beta_2, \ldots, \beta_{d_1}, 0,0,\ldots,0\right)$, d-$d_1$>0.
- If a lasso is trained, w.r.t. prediction accuracy, then it selects the true model with probability c<1, which depends on the error variance and d_1 but not on n.
- Cand random forest, buruta, etc. do any better?

Perils (cont.): A Gene Expression example

- Take another example from Effron (2020): A study involved n = 102 men, 52 cancer patients and 50 normal controls. Each man's genetic expression levels were measured on a panel of p = 6033 genes.
- Aim is to predict predicting normal or cancer from a man's microarray measurements.
- Random Forest, can easily be used to compute variable importance. But hat happens when we remove important genes from the data and refit random forest?

Prediction of error of Random Forest after removal of top-most important variables

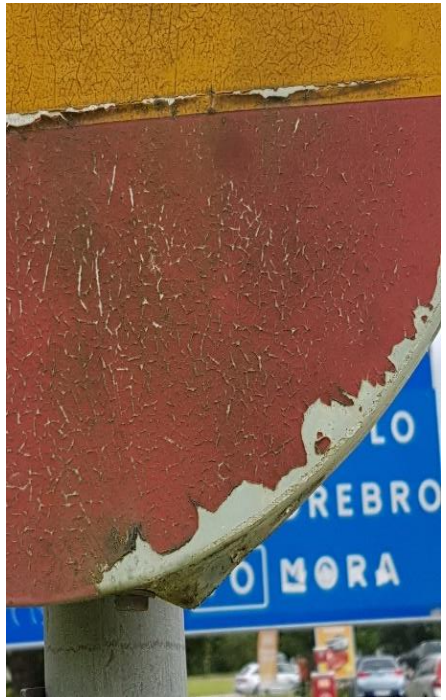
# removed	0	1	5	10	20	40	80	160	348
# errors	1	0	3	1	1	2	2	2	0

Source: Effron, 2020

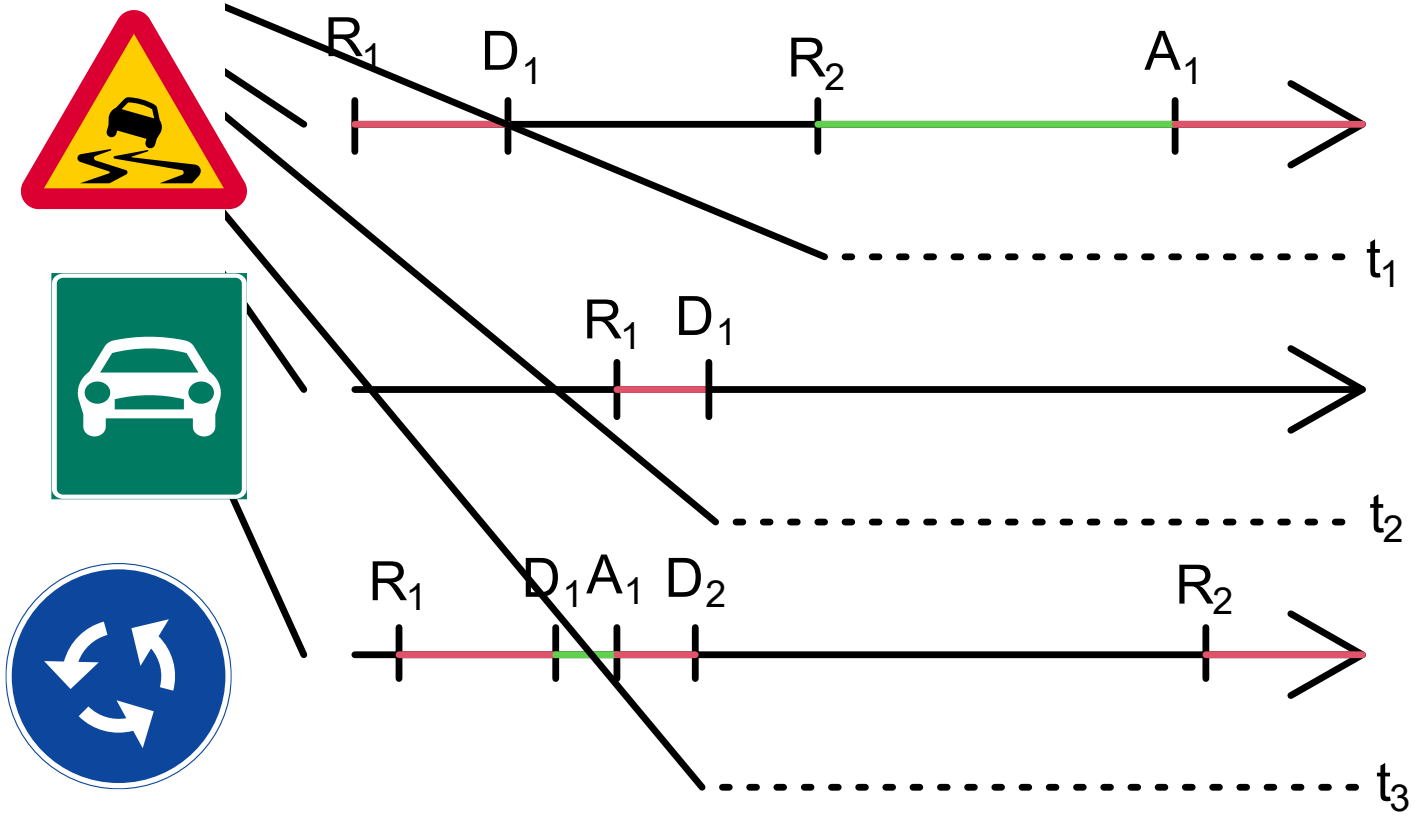
Estimation of service lifetime of the traffic sign

- The transportation agency in Sweden (Trafikverket) wishes to establish an optimal maintenance schedule for the traffic signs.
- Data pertaining to the color and retro reflectivity of traffic sign were collected using a survey conducted with a hand-held spectrometer.
- Based on the spectrometric reading, the status of the traffic sign were determined, in accordance with the European standard, into two classes (acceptable, and not acceptable).
- The date of manufacturing of the signs, along with geographical location specific data were also collected.
- The focus of the first task is to estimate the length of service life of the traffic sign, using the survey data.

Field data collection



The observational data (on partial lifetime)



Legend:
R: Real event (lifetime),
D: Replacement delay,
A: Time to accident,
t: Observed life in service

Theoretical model

Let us define different hazard functions for,

R (real lifetime) : $h_R(t) = \gamma\lambda e^{X\beta} t^{\gamma-1}$ (1)

A (accident) : $h_A(t) = \theta$ (2)

D (replacement delay): $h_D(t) = \tau$ (3)

Using (1) and (2), for the observed service lifetime, we have

O (min{R,A}): $h_O(t) = h_A(t) + h_R(t) = \theta + \gamma\lambda e^{X\beta} t^{\gamma-1}$ (4)

Using (3) and (4), we have: $S_{O+D}(t) = \int_0^t \tau \exp[-\tau t] \left(\int_0^\infty \left(\left(\gamma\lambda e^{\boldsymbol{x\beta}} u^{\gamma-1} + \theta \right) \exp\left[-\left(\theta u + \lambda u^\gamma e^{\boldsymbol{x\beta}} \right) \right] \right) \exp[\tau u]\, du \right) dt$

Decoding the problem with logistic distribution

- We observe whether the object is in functional (State, S =1) or in dysfunctional (S =0) state, and time spent in that state.
- The observed response pair (S, T), along with covariates, X.
- Using properties of the renewal process, at steady state, we have

$$P(S=1)=\frac{E(O)}{E(O)+E(D)},\ P(S=0)=\frac{E(D)}{E(O)+E(D)}$$

- Here, $E(D)=\frac{1}{\tau}$, and $E(O)=\sum_{j=0}^{\infty}\frac{\left(-\lambda e^{\boldsymbol{x\beta}}\right)^{j}}{j!\theta^{1+j\gamma}}\left[\Gamma(2+j\gamma)+\frac{\gamma\lambda e^{\boldsymbol{x\beta}}\Gamma(1+(j+1)\gamma)}{\theta^{\gamma}}\right]$
- This implies that if one fits a logistic regression model with this data, it will be impossible to interpret the results.

Summarizing the examples

- The subject of Statistics is going through a period of self rediscovery, and it did that many times in the past.
- Exiting empirical evidences brought about a lot of enthusiasm about the potentials of data driven solutions.
- Statisticians should come forward to challenge the current practice of data analysis and establish their theoretical basis.
- Core theoretical knowledge of Statistics still matters for solving complex real-world problems.

References

- Agresti, A., & Meng, X. (2013), Strength in Numbers: …, Springer
- d'Errico, F., Doyon, L., Colagé, I., et al.. (2018), From number sense to number symbols. An archaeological perspective. Phil. Trans. Royal Society B: Biolog. Scie., 373 (1740)
- Cleveland, W.S. (2001), Data Science: An action plan for expanding the technical areas of the field of Statistics, Intl. Statist. Rev., 69.
- Cuadrado-Gallego, J.J. and Demchenko, Y. (2020), The Data Science Framework A View from the EDISON Project, Springer, Cham.
- Donoho, D. (2017), 50 years of Data Science, J. Comp. Graph. Statist., 26(4).
- Efron, B. (2020), Prediction, estimation, and attribution, J. Amer. Statist. Assoc., 115(530).
- Leng, C., Lin, Y., & Wahba, G. (2006), A note on the lasso and related procedures in model selection, Statis. Sinica, 16(4), 1273-1284 .
- Meitzen, A. (1886), History, Theory, and Technique of Statistics., R. P. Falkner (trans., 1891), Amer. Acad. Social & Political Scie..
- Sjöström, O. (2002). Svensk statistikhistoria – en undanskymd kritisk tradition.