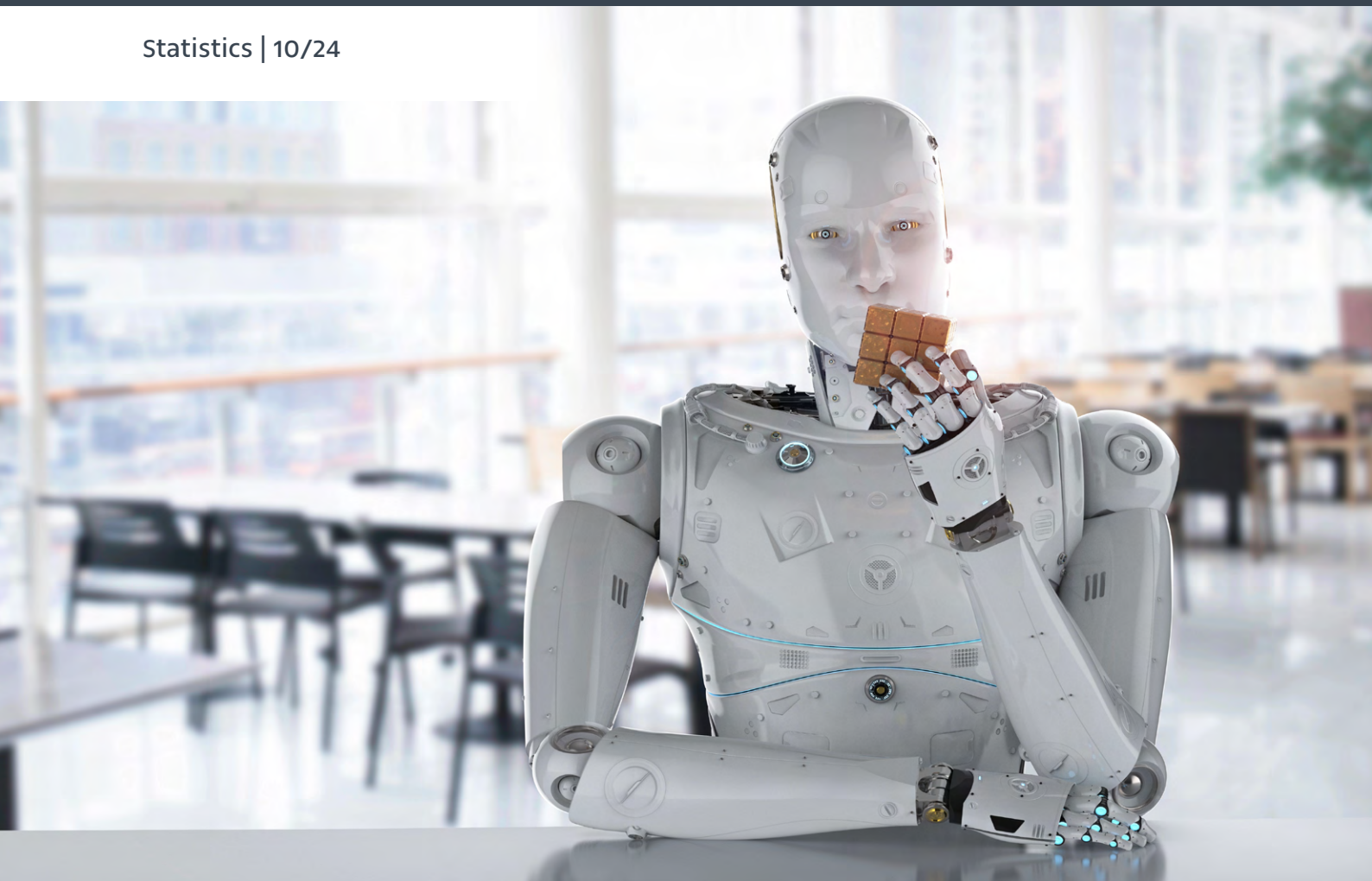
A Tripartite View on the Role of AI in Modern Analytics

Ron S. Kenett

SAMUEL Neaman Institute FOR NATIONAL POLICY RESEARCH

Statistics | 10/24



SAMUEL
Neaman Institute
FOR NATIONAL POLICY RESEARCH

A Tripartite View on the Role of AI in Modern Analytics

Ron S. Kenett

Senior Research Fellow, Samuel Neaman Institute, Technion

Chairman, KPA Ltd, Chairman, Data Science Society, AEAI

Professor, University of Turin, Italy

October 2024

Contents

Abstract

Artificial intelligence (AI) and machine learning (ML) are gaining presence in all aspects of research and data analysis. Strictly speaking, ML is a subfield of AI about the algorithms and statistical tools that allow computer systems to perform specific tasks without explicit instructions. One side effect of this evolution is an expanded interest in data driven studies including topics such as data integration, alternative scenario analysis and predictive analytics. This paper is a retrospective view of experience gained by applying statistics and analytics to a wide range of problems, with an emphasis on the past few years. It aims to show how AI and ML merge with statistics in a powerful modern analytic environment. The paper has three parts. Part I is about AI and statistics practice, Part II is about AI in Medicine and Part III is about AI in survey design and analysis. If needed, references provide more details.

Part I: On AI and Statistics[1]

[1] Adapted from a blog in https://errorstatistics.com/2024/07/11/guest-post-ron-kenett-whats-happening-in-statistical-practice-since-the-abandon-statistical-significance-call-5-years-ago/

An important impact on the current practice of statistics is the merging of empirical predictive analytics with probability-based modelling. In many applications, but not in all, one has access to massive data coming from sensor technologies and unstructured formats such as text and images. In these cases, methods to fit and assess a model are based on splitting the data into a training set and a validation set. The training data is used to fit a model, the validation set, to evaluate it. If the fit of a model to the training data is very good but the fit to the validation set is low, we experience what is labeled “overfitting”. Overfitting reduces the ability to generalize the model to future implementations and implies poor predictive performance. This approach is different from the classical statistical analysis relying on probability models and hypothesis testing.

R. A Fisher, in his fundamental paper “On the Mathematical Foundations of Theoretical Statistics.”, stated that “the object of statistical method is the reduction of data.” He then identified “three problems which arise in the reduction of data.”: Problem 1: *Specification*—choosing the right mathematical model for a population. Problem 2: *Estimation*—methods

to calculate estimates, from a sample, and Problem 3: *Distribution*—properties of estimators derived from samples. The predictive analytic methods outlined in the first paragraph change all this. The *Specification, Estimation, Distribution* trio is replaced, with computer intensive methods including cross validation, bootstrapping and simulations. The models used in this context are supervised or unsupervised, with transfer learning, active learning labelling, zero shot and few shot learning.

In this context of new problems and new methods, we still face the fundamental issue of collecting data with relevance to the problem at hand, what Colin Mallows called the "zeroth problem." For example, splitting of the data into training and validation must be consistent with the data generation process. In an industrial batch processing plant we will want to split the data by entire batches and not by individual measurements. Principles for achieving this are formulated in an approach titled befitting cross validation (BCV), see Kenett et al (2022, https://xwdeng80.github.io/BCV2022.pdf). Data scientists developing predictive models need to be aware of BCV.

The combination of computer intensive methods and classical statistical methods (Bayesian or frequentists) is a challenge for current work in data analysis. In many companies, Statistics groups have been replaced by Data Science groups. Some universities do not even involve statisticians in their data science programs. This creates a risk where past experience gained by Statisticians is lost and data analysis mistakes are repeated. Blind applications of AI involve similar risks and the statistical evaluation of AI is crucial for the AI trustworthy (Lian et al, 2021, Hong et al, 2023).

With this perspective, let me share my views on what's happening in statistical practice and AI, with consequences to data analysis.

The American Statistical Association organised on October 2017, in Bethesda Maryland, a *symposium on statistical inference* (SSI). I summarized what happened there in a blog titled "to p or not to p". See: https://blogisbis.wordpress.com/2017/10/24/to-p-or-not-to-p-my-thoughts-on-the-asa-symposium-on-statistical-inference/ .

Today, there are many advances in data analysis coming from Statistics and AI. Some examples include, selection bias (Benjamini, 2019), fairness in analytic models (Plecko and Bareinboim, 2023) and information quality. Some questions and answers on how to address current challenges in data analysis were presented in this seminar , See also here. A summary of these points is listed below:

1. How should we practice Statistics and analytics? Embrace a life cycle perspective, from problem elicitation to generalization, operationalization and communication of findings.

2. How should we teach Statistics and analytic courses? Engage the students with real life projects and dedicate time to the conceptual understanding of Statistical methods and thinking, AI and ML.

3. What are research areas for statistics and analytics to focus on? Areas at the interface of Statistics with Machine Learning/Artificial Intelligence/Computer Science/Data Science. Specifically, how can AI be used in Statistical analysis?

4. How do we initiate synergistic collaborations between disciplines? By direct communication. We need forums enabling such exchanges.

5. What is the role of professional organizations in this transformation? Professional organizations have a unique responsibility to foster discussion and provide an opportunity for contrarian views to be expressed. This includes professional Statistics organisations such as ENBIS and INFORMS.

6. How should lifelong learning be implemented to update the skills of working data analysts? Adult education is posing a different challenge from the one faced in regular academia. In that context, simulation-based education and online training material are excellent options.

In summary, the positioning of analytics is at a peak. A conference on “the foundations of applied statistics” was held at the Samuel Neaman Institute, Technion, Israel on April 2024. Presenters addressed various aspects of applied statistics including philosophical methods, historical examples and designing experiments for generalizability of findings. For slides and a video recording see https://www.neaman.org.il/en/events/on-the-foundations-of-applied-statistics-april-2024/ .

In the future, Statistics and AI will bring up new innovative and exciting developments. These need to be discussed and properly managed.

Part II: On AI and Medicine[2]

[2] Based on a talk prepared for the “Symposium sur l'A.I.” at the Clinique Notre-Dame de Grâce (CNDG) in Gosselies, Belgium.

Doctors and nurses often ask: "what could we do as individuals to be prepared in our practice with the upcoming of the AI?". As a first step, my response is: “be aware of the potential in AI and ML”. This section provides an overview of applications of AI and ML in

medicine. Examples include blood pressure monitoring, time use epidemiology, medical imaging, animal studies, integrated data analysis during COVID19, processing of lipoaspirate for generative applications, immunotherapy for children with peanut allergy and clinical determinants of response to HBI0101 (CART) therapy in myeloma. In these examples, sensor data, statistical models, AI and ML provide an analytic perspective enhancing our ability to understand and predict clinical outcomes. The section covers application examples of: Data in medicine, cross validation, association rules, integrated analysis, functional metanalysis, selective inference, decision trees and latent class analysis. A generic plan for action will be offered at the end of the section listing action items and directions to take.

We start with a quick review of the last 350 years. Sir William Petty (1623-1687) is considered to be the first modern epidemiologist. He was much influenced by Francis Bacon, was in touch with Renee Descartes, and held the conviction that mathematics and the senses must be the basis of all rational sciences. As a physician he related his knowledge of health and disease to social epidemiology statistics. Florence Nightingale (1820 – 1910) came to prominence while serving as a nurse during the Crimean War, where she organised care for wounded soldiers at Constantinople. She proposed novel graphical forms to convince decision makers that soldier deaths were mostly due to sanitary conditions rather than battle wounds. She is famous for usage of the polar area diagram which is equivalent to a modern circular histogram. This diagram is still regularly used in data visualisation. John Snow (1813 – 1858) was a physician famous for his work in tracing the source of a cholera outbreak in London's Soho, which he identified as a particular public water pump using location plots marking homes of affected inhabitants. Building on the work of Hume and Popper, Bradford Hill (1897 – 1991) proposed in 1965 a set of nine criteria to provide epidemiologic evidence of a causal relationship between a presumed cause and an observed effect. Researchers applied Hill’s criteria in many areas of epidemiology such as connections between exposures to molds and infant pulmonary hemorrhage, ultraviolet B radiation and cancer, vitamin D and neonatal outcomes, alcohol and cardiovascular disease outcomes, infections and risk of stroke and foods and nutrients related to cardiovascular disease and diabetes. Our last stop on this journey is Sir David Cox (1924 – 2022), a British statistician and educator. Cox contributions to the field of statistics include logistic regression, the Cox proportional hazards model and the Cox stochastic process. His impact on statistics and clinical research is immeasurable. These giants laid foundations for data driven clinical research and modern epidemiology.

In recent years a major change occurred with the advent of computer-based analytics, massive data and sensor technology. Some numbers quoted mention that a person generates 1000 Terabyte of health data during his life and that every 73 days, heath data is doubling. To put this change in context we first describe two studies where data is collected

through various devices. This is followed by other studies with examples implementing analytics. The first study is about casual blood tests performed by nurses and data collected through Holter ambulatory monitoring devices (Weiss et al 2016). The study reports that elevated systolic blood pressure, recorded by 24-hour blood pressure monitoring on the first day of acute stroke of elderly patients, is associated with an unfavourable outcome. Clinically, elevated blood pressure during an acute ischemic stroke might be advantageous by improving cerebral perfusion to the ischemic tissue or detrimental by exacerbating oedema and haemorrhagic transformation of the ischemic tissue. It is therefore unclear whether high blood pressure should be treated in acute ischemic stroke. The study assessed the impact of the change in systolic blood pressure levels in elderly patients measured with a monitoring device during the first week of stroke on short-term functional status and long-term mortality. In the elderly patients with acute stroke, the change in systolic blood pressure had no effect on short-term improvement in functional outcome. The study confirmed findings in the literature that elevated admission, as assessed by 24 hours monitoring, are associated with poor short- and long-term outcomes. It provided an evaluation of monitoring versus casual testing and a reproducibility of results presented in the literature. Reproducibility of outcomes is at the core of scientific research. It is focused on the research claim and should be distinguished from replicability which aims at replicating effect estimates (Kenett and Shmueli, 2015).

The second study is about time use epidemiology with data collected by Fitbit bracelets given to children (Dumuid et al, 2022). The Fitbit data is analysed to determine the time allocation by type of activity. This data is then analysed in terms of health-related measures and a software application for planning a day's activity, is provided in the paper. The analysis evaluates alternative time use scenarios, for example: more play time and less sleep. The blood pressure measurement in the first study was diagnostic, the time use model in the second study is prognostic. Two different uses of sensor data.

The next two studies are on modelling. Zinger et al (2023) is based on a metanalysis of papers on mechanical processing of lipoaspirate used as a source for regenerative cells in stem cell transplants used to manage orthopedic pathologies. A review of the literature listed different techniques of lipoaspirate extraction. Data from the articles were integrated to create a statistical-based predictive model for multiple input variables, what one might call a functional metanalysis design of experiment. The response variable is the normalized yield of nucleated cells per mL of starting lipoaspirate obtained through the mechanical processing. Sixteen input factors were extracted from reported studies and used in fitting a random effect model corresponding to the different studies involved. Kfir-Erenfeld et al (2024) is analysing a is single-arm, open-label study evaluating locally produced BCMA-CART HBI0101 and conducted at hospital Department of Bone Marrow Transplantation and Cancer

Immunotherapy. HBI0101 is an academic chimeric antigen receptor T-cell (CART)–targeted to B-cell maturation antigen (BCMA) for the treatment of relapsed and refractory multiple myeloma and light chain amyloidosis. A Cox proportional hazard model on patient progression-free survival and CART-related characteristics revealed that high-risk cytogenetic, extramedullary disease, and increased number of effector-memory T cells in CART products were independently associated with inferior PFS. The lipoaspirate extraction metanalysis is applying a design of experiments methodology with random effects, the CART Cox proportional hazard model is controlling for selective inference. Design of experiments is about designed interventions used to model the effect of input factors on responses. Selective inference is affecting studies where decisions on modelling are done after seeing the data. This effect is biasing results and needs to be accounted for. Both the design of experiments and the Cox proportional hazard model are widely used in clinical research.

The next studies we refer to apply methods less well known in clinical research. Predictive analytics are based on models validated by splitting data into training sets and validation sets. The models are developed (learned) with the training set and validated with the validation data, where the model prediction is compared to the outcome in the data. In splitting the data, one needs to account for the data generation process. Kenett et al (2022) propose an approach to data splitting labelled befitting cross validation (BCV) that matches the process generation to the splitting approach. In patient medical records one can focus on individual lab tests, on patient visits or on patients. Each of these options leads to different cross data splitting approaches. A similar example occurs in reports of drug product side effects. One can consider individual reports, reports by patients or drug product batches. Other methods mentioned below in the context of clinical research, are decision trees and latent class analysis. Yahia et al (2022), Machnes-Maayan et al (2022) and Cohen et al (2024) study children with allergies. In these studies, decision trees and latent class analysis are used. Decision trees classify cases (allergic children) and set up data driven thresholds. Latent class analysis identifies unobserved latent variables representing data clusters. Decision trees are producing clinical protocols for treating patients, latent class analysis is used in identifying patient profiles. Finally, Kenett et al (2022) apply Bayesian networks and structural equation models to integrate health related data and mobility data, from Italy and Israel, during the COVID pandemic. The structural equation model is testing the significance of effects. The Bayesian network permits to assess alternative scenarios corresponding to non-pharmaceutical interventions. As a sample scenario, the paper considers the effect of closing airports on the number of patients admitted to hospital ICUs. In these reviewed studies, we applied a wide range of methods including design of experiments, Cox proportional hazard models, decision trees and latent class analysis.

Details are available in the publications listed below. A skeleton plan to meet challenges involved in such applications includes:

1. Problem elicitation – formulate the goal of a study in a wide sense
2. Data sources mapping – identify sources of relevant data
3. Data integration – can be done for individual or aggregated data
4. Data clean up – data preprocessing is always required
5. Descriptive and visualization tools – a good place to start
6. Analytic competencies – expose clinical researchers to modern analytics
7. Supervised and unsupervised learning – lots of options here
8. Operationalization of findings – this requires interaction between clinicians and analysts
9. Communication of findings – consider how to address several stakeholders

To meet such challenges, one needs motivation and competencies. Organizations meeting the challenges above rise in analytic maturity. The examples in Part II are designed to facilitate this.

Part III: On AI and Surveys[3]

[3] Based on a talk prepared for the ISDSA symposium on Surveys commemorating the work of Professor Camil Fuchs, Tel Aviv University.

This section provides an overview of applications of AI and ML in the design and analysis of surveys. It covers topics relevant to survey data analysis such as generalizability of findings, applications of Bayesian networks, decision trees and text analytics. All these topics are introduced with references to the literature where details can be found. The section is a structured review providing a wide-angle perspective on AI and Survey design and analysis.

In the preface to *The Economic Control of Quality of manufactured product* by W. Shewhart (1931), W. Edwards Deming wrote: “Tests of variables that affect a process are useful only if they predict what will happen if this or that variable is increased or decreased. Statistical theory, as taught in the books, is valid and leads to operationally verifiable tests and criteria for an enumerative study. Not so with an analytic problem, as the conditions of the

experiment will not be duplicated in the next trial. Unfortunately, most problems in industry are analytic." This distinction between enumerative and analytic studies was apparently conceived in an industrial setting. It does, however, apply also in survey design and analysis (Deming, 1953). Enumerative studies are the classical realm of sample surveys. The study goal is a population frame, the practical goal is a sampling frame which overlaps the population frame to a large extent. For example, we want to study the preferences of inhabitants in an urban neighborhood, and we have access to a registry managed by the municipality. A sample is drawn from this registry in order to infer population parameters, for example preferences regarding a new municipal park: Should it have a bike lane, or not. This basic inference can be expanded if one considers generalizability of findings (Pearl, 2015, Kenett and Shmueli, 2016). We discuss all these issues below. Specifically, we introduce the application of Bayesian networks to the analysis of survey data. We also show how Bayesian networks are used for integrating data sources. Bayesian networks are typically classified as an AI or ML tool. We show how using BNs moves survey data analysis to the analytic domain enabling the evaluation of alternative scenarios. An example of this, in the COVID pandemic, is provided Kenett et al (2022). Other tools from AI, in the context of surveys, include decision trees and text analytics. We start with the Deming distinction between enumerative and analytic studies and then review respondent driven sampling, a promising method for the design of surveys in the modern ecosystem of linked communities.

In a general framework of information quality, Kenett and Shmueli (2016) list eight dimensions including *Operationalization of Findings.* This emphasizes a consideration of statistical analysis as a basis of action. Surveys designed to estimate properties of a sampling frame, on the basis of sample data, are enumerative. In this context one distinguishes between probability samples where random samples are drawn from a sampling frame list, from quota samples derived from panels where recruitment is up to the participants. In probability samples one can apply probability models to assess uncertainty in estimates. A long practice approach to random sampling is to draw a sample from a listing of a frame and collect questionnaire responses from the sample. Not everyone agrees to respond, creating a non-response effect. With availability of emails of customers, companies carry out a census by approaching all customers with an email invitation to fill in a questionnaire. This also results in non-responses. In random sampling one can move on to the next person on the list. With an attempted census one can only send reminders and issue follow-up calls. If the non-responding customers have specific characteristics this can severely bias the population estimates. The M test proposed in Fuchs and Kenett (1980) is used to assess representativeness of responses vis a vis the overall targeted population. It compares the actual responses to expected responses, correcting for multiplicity of classifications using a Bonferroni correction. In quota samples one can weigh results but not

assess uncertainty. Analytic studies require counterfactual intervention such as in alternative scenario analysis. We cover this option later in the section with Bayesian networks. At this point we just emphasize the distinction between enumerative and analytic studies. A modern discussion of this distinction is provided by Shmueli (2010) where models aiming at presenting causality explanations are put in contrast to models used in predictive analytics. The use of AI and ML in survey data analysis enables analytics modeling, beyond the classical enumerative approach to survey data analysis such as in Cochran (1977). Part III focuses on these options.

In many cases one has difficulty identifying a sampling frame. On the other hand, communities are commonly formed by social media platforms enabling easy access to potential survey respondents. With the gradual drop in responses to telephone-based surveys, Internet panels have become the prime platforms for surveys (Kenett et al, 2018). Panels are amenable to quota sampling but not to random sampling from a survey frame. A partial understanding of possible bias in the data collected through panels requires details on entry and exit criteria of panel participants.

Heckathorn and Cameron (2017) review network sampling methods that provide coverage of target populations without sampling frames. An example is *snowball sampling*, where a survey begins with a convenience sample of initial subjects who serve as seeds. Sampling then proceeds through network linkages, first from the seeds to the first wave, then from the first to the second wave, and so forth as the sample expands from wave to wave in the manner of a snowball growing as it rolls down a hill. Sampling stops when the target sample size has been attained. The major limitation of this method is that neither the selection of the initial subjects nor the selection of the subsequent waves is random, so that statistical inference from it to the sampling frame is impossible. A similar type of design is link-tracing designs. This approach assumes that everyone in the country could hypothetically be reached by the sixth wave of a maximally expansive link-tracing. Multiplicity sampling draws on respondents' knowledge of their own networks so that respondents are asked about events among those in their personal networks. Respondent-driven sampling (RDS) is a form of network sampling that combines both multiplicity and link-tracing methods. It provides mathematical adjustments that can be applied to a sample to compensate for biases resulting from the network structures that affect sampling, thereby yielding a form of probability sample. Markov modelling of the peer recruitment process shows that bias from the convenience sample of initial subjects in link-tracing is progressively attenuated as the sample is expanded wave by wave. RDS uses data from peer recruitments to estimate the probability of recruitment across groups. These probabilities serve as the transition probabilities of the Markov model. Equilibrium is independent of the starting point, that is, independent of the convenience sample of seeds from which it began. Furthermore, the

analysis shows that bias from the seeds is reduced at a geometric rather than an arithmetic rate, a feature that accelerates the reduction of bias. There are two main approaches for RDS sampling variance estimation: bootstrap and analytic approaches. Research on RDS is still ongoing but the approach is able to handle surveys on communities accessible through the media, without sampling frames. AI and ML, applied to social media networks open up strong opportunities for RDS applications.

After discussing the design of surveys through network sampling methods we move on to methods for analysing survey data. This is where most of the impact of AI and ML is/should be felt. We start by discussing Bayesian networks.

Bayesian networks (BNs) implement a graphical model structure known as a directed acyclic graph (DAG) that is popular in statistics, machine learning and artificial intelligence. BNs enable an effective representation and computation of the joint probability distribution over a set of random variables (Pearl, 1985). The structure of a DAG is defined by two sets: the set of nodes and the set of directed arcs. The nodes represent random variables and are drawn as circles labelled by the variable names. The arcs represent links among the variables and are represented by arrows between nodes. In particular, an arc from node Xi to node Xj represents a relation between the corresponding variables. Thus, an arrow indicates that a value taken by variable Xj depends on the value taken by variable Xi. Node Xi is then referred to as a 'parent' of Xj and, similarly, Xj is referred to as the 'child' of Xi. This property is used to reduce the number of parameters that are required to characterize the joint probability distribution of the variables. This reduction provides an efficient way to compute the posterior probabilities given the evidence present in the data. In addition to the DAG structure, which is often considered as the qualitative part of the model, a BN includes quantitative parameters. These parameters are described by applying the Markov property, where the conditional probability distribution at each node depends only on its parents. For discrete random variables, this conditional probability is represented by a table, listing the local probability that a child node takes on each of the feasible values – for each combination of the values of its parents. The joint distribution of a collection of variables is determined uniquely by these local conditional probability tables.

BNs can be specified by expert knowledge or learned from data, or in combinations of both. In learning the network structure, one can include whitelists of forced causality links imposed by expert opinion and blacklists of links that are not to be included in the network. The parameters of the local distributions are learned from data, priors elicited from experts, or both. Learning the graph structure of a BN requires a scoring function and a search strategy. Common scoring functions include the posterior probability of the structure given the training data, the Bayesian information criterion (BIC) or Akaike information criterion

(AIC). When fitting models, adding parameters increases the likelihood, which may result in over-fitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model, with the penalty term being larger in BIC than in AIC. A partial list of structure learning algorithms includes Hill-Climbing with score functions BIC and AIC, Grow-Shrink, Incremental Association, Fast Incremental Association, Interleaved Incremental Association, hybrid algorithms and Phase Restricted Maximization.

To fully specify a BN, and thus represent joint probability distributions, it is necessary to specify for each node X the probability distribution for X conditional upon X's parents. The distribution of X, conditional upon its parents, may have any form with or without constraints. These conditional distributions include parameters which are often unknown and must be estimated from data, for example using maximum likelihood. Direct maximization of the likelihood (or of the posterior probability) is usually based on the expectation-maximization (E-M) algorithm which alternates computing expected values of the unobserved variables conditional on observed data, with maximizing the complete likelihood assuming that previously computed expected values are correct. Under mild regularity conditions this process converges to maximum likelihood (or maximum posterior) values of parameters

Discrete BNs are particularly well suited to analyze anchor-based surveys (Kenett and Salini, 2009). One of the benefits of BNs is that they enable the assessment of alternative scenarios using conditioning on parent variables (Kenett and Salini, 2012). This capability provides decision makers with a decision support system based on results of surveys. For example, in a survey on customer satisfaction a company can determine the effect of improving satisfaction from technical support on overall satisfaction and level of recommendation. Another application of BNS is the integration of data coming from different surveys. Dalla Valle and Kenett (2018) combine with BNs data from face-to-face interviews at San Francisco airport with data gathered from a website specializing in collecting customer satisfaction data from experience at airports. By fitting BNs of both sets of data and calibrating them one gets a wider perspective. The calibration in the case of the San Francisco airport was using survey data on specific experience at the airport from security screening procedures. Integrating data sources enhances the impact of survey data. This can be based on combining surveys or on combining with surveys operational data.

Another type of integration is not on the data itself but on the outcomes of analysis using various models. Kenett and Salini (2011) analyze the same survey data with four models who provide complementary advantages. They use the information quality framework of Kenett and Shmueli (2016) to compare the models and eventually propose an analysis combining these models for enhanced information quality.

Survey data are collected in a certain context at a certain point in time. Generalizability, sometimes called transportability, is transferring knowledge learned in a certain survey to other situations. Pearl (2015) shows how BNs can be used to conduct this transfer. It leverages the conditioning ability in BNs to adapt results. An example is a survey conducted in Los Angeles whose results are applied to New York with an older population. In analyzing Bayesian networks on can consider a basic association level or more complex causality links (see Kenett, 2019).

In summary, BNs are offering significant capabilities to survey data analysis. They provide a natural approach to the analysis of Likert scale discrete data, permit the evaluation of alternative scenarios, enable integration of data at both the individual and aggregated level and, finally provide a structure for generalization of findings. An additional advantage of BNs is that they anonymize responders. Companies running employee surveys are specially sensitized to anonymity of responses, a BN will do that.

The application of BNs can be considered an application of AI and ML. We next present an analysis of survey data with decision trees, a popular AI method. Assume again a customer satisfaction survey. We can code the overall satisfaction question into binary options: satisfied, not satisfied. To characterize these two groups, we can use decision trees, random forests or boosted trees. This will highlight the determinants of customer satisfaction with thresholds. If the survey involves 100-200 responses, we apply an exploratory decision tree to all the data. If we have over 1000 responses, we can split it into training and validation subsets and derive a predictive model. One application of this analysis is labeling customers as potential churn and applying to them churn prevention procedures. If they contact the call center, they get special treatment with special offers and speedy response.

Surveys typically include an option to provide written comments. Text analytics using Latent Class Analysis and Latent Semantic Analysis using partial singular value decomposition provide insights on topics discussed in the text (Liang et al, 2023, Gahre-Daghi, 2024). Text analysis can also be conducted using Large Language Models (LLM). For example, Shahin et al (2024) apply ChatGPT 3.5 to analyse comments generated by customers. The drawback of applying LLMs like ChatGPT is that the responses need to be verified before being used so that the research emphasis moves from efforts to generate information to efforts to ensure the quality of information (Kenett, 2024).

In reviewing applications of AI and ML to survey data analysis we conclude with a mention of comments typically included in survey questionnaires (Kenett et al, 2023). Text analytics can identify topics discussed with a comparison of topics of satisfied and dissatisfied customers. A further analysis is based on weights of terms leading to sentiment analysis. Finally, a general comment worth accounting for before the study is conducted is that in conducting text analysis one needs to ensure active of passive informed consent.

References

Cochran, W. G. (1977). *Sampling Techniques* Third Edition John Wiley & Sons Inc. *New York*.

Cohen, C. G., Levy, Y., Manasherova, E., Agmon-Levin, N., Kenett, R. S., Jean-claude, B. J., ... & Kidon, M. I. (2024). Peanut Allergen Characterization and Allergenicity Throughout Development. *Frontiers in Allergy*, 5, 1395834.

Dalla Valle, L., & Kenett, R. (2018). Social media big data integration: A new approach based on calibration. *Expert Systems with Applications*, 111, 76-90.

Deming, W. E. (1953). On the distinction between enumerative and analytic surveys. *Journal of the American Statistical Association*, *48*(262), 244-255.

Dumuid, D., Olds, T., Wake, M., Rasmussen, C. L., Pedišić, Ž., Hughes, J. H., ... & Stanford, T. (2022). Your best day: An interactive app to translate how time reallocations within a 24-hour day are associated with health measures. *Plos one*, *17*(9), e0272343.

Fuchs, C., & Kenett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *Journal of the American Statistical Association*, *75*(370), 395-398.

Gahre-Daghi, N. (2024) The analysis of customer comments in insurance, thesis in Quantitative Finance and Insurance, Supervisor: Luciano, e. And Kenett, R.S., University of Turin, Italy

Halabi, A., Kenett, R. S., & Sacerdote, L. (2017). Using dynamic Bayesian networks to model technical risk management efficiency. *Quality and Reliability Engineering International*, *33*(6), 1179-1196.

Harel, A., Kenett, R. S., & Ruggeri, F. (2008). Decision support for user interface design: usability diagnosis by time analysis of the user activity. In *32nd Annual IEEE International Computer Software and Applications Conference* (pp. 836-840). IEEE.

Heckathorn, D. D., & Cameron, C. J. (2017). Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annual review of sociology*, *43*(1), 101-119.

Hong, Y., Lian, J., Xu, L., Min, J., Wang, Y., Freeman, L. J., & Deng, X. (2023). Statistical perspectives on reliability of artificial intelligence systems. *Quality Engineering*, *35*(1), 56-78.

Kenett, R. S. (2016). On generating high InfoQ with Bayesian networks. *Quality Technology & Quantitative Management*, *13*(3), 309-332.

Kenett, R. S. (2019) Cause-and-Effect Diagrams, Wiley StatsRef Encyclopedia, John Wiley and Sons.

Kenett, R. S. (2024) Multivariate techniques for analyzing and presenting official statistics indicators, *Statistica Applicata - Italian Journal of Applied Statistics*, Vol. 36, Number 1.

Kenett, R. S., & Bortman, J. (2022). The digital twin in Industry 4.0: A wide-angle perspective. *Quality and Reliability Engineering International*, *38*(3), 1357-1366.

Kenett, R. S., & Maggino, F. (2021). Techniques for analyzing and presenting official statistics indicators. *Statistical Journal of the IAOS*, *37*(2), 541-552.

Kenett, R. S., & Salini, S. (2009) New Frontiers: Bayesian networks give insight into survey-data analysis, Quality Progress, 42(8):30-36

Kenett, R. S., & Salini, S. (2011). Modern analysis of customer satisfaction surveys: comparison of models and integrated analysis. *Applied Stochastic Models in Business and Industry*, *27*(5), 465-475.

Kenett, R. S., & Shmueli, G. (2016*). Information quality: The potential of data and analytics to generate knowledge*. John Wiley & Sons.

Kenett, R. S., Gal, R., Adres, E., Ali, N., & Glickman, H. (2023). The Israeli society common denominator: a text analytic study. *Israel Affairs*, 29(3), 669-679.

Kenett, R. S., Gotwalt, C., & Poggi, J. M. (2024). An analytic journey in an industrial classification problem: How to use models to sharpen your questions. *Quality and Reliability Engineering International*, 40(2), 803-818.

Kenett, R. S., Gotwalt, C., Freeman, L., & Deng, X. (2022). Self-supervised cross validation using data generation structure. *Applied Stochastic Models in Business and Industry*, *38*(5), 750-765.

Kenett, R. S., Manzi, G., Rapaport, C., & Salini, S. (2022). Integrated analysis of behavioural and health COVID-19 data combining Bayesian networks and structural equation models. *International Journal of Environmental Research and Public Health*, 19(8), 4859.

Kenett, R. S., Pfeffermann, D., & Steinberg, D. M. (2018). Election polls—a survey, a critique, and proposals. *Annual Review of Statistics and Its Application*, 5(1), 1-24.

Kenett, R. S., Zacks, S., & Gedeck, P. (2022). *Modern Statistics: A Computer-Based Approach with Python*. Cham: Springer International Publishing.

Kenett, R. S., Zacks, S., & Gedeck, P. (2023). *Industrial Statistics: A Computer-Based Approach with Python*. Cham: Springer International Publishing.

Kenett, R., & Salini, S. (2012). *Modern Analysis of Customer Surveys*. John Wiley and Sons.

Kenett, R., Shmueli, G. (2015) Clarifying the terminology that describes scientific reproducibility. *Nat Methods* **12**, 699. https://doi.org/10.1038/nmeth.3489

Kenett, R.S. (2024) Some notes on ChatGPT and information quality, SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4760077

Kfir-Erenfeld, S., Asherie, N., Lebel, E., Vainstein, V., Assayag, M., Dubnikov Sharon, T., ... & Stepensky, P. (2024). Clinical evaluation and determinants of response to HBI0101 (BCMA CART) therapy in relapsed/refractory multiple myeloma. *Blood Advances*, *8*(15), 4077-4088.

Lian, J., Freeman, L., Hong, Y., and Deng, X. (2021). Robustness with Respect to Class Imbalance in Artificial Intelligence Classification Algorithms, *Journal of Quality Technology*, 53(5), 505-525.

Liang, Q., Ranganathan, S., Wang, K., & Deng, X. (2023). JST-RR model: joint modeling of ratings and reviews in sentiment-topic prediction. *Technometrics*, *65*(1), 57-69.

Machnes-Maayan, D., Yahia, S. H., Frizinsky, S., Maoz-Segal, R., Offengenden, I., Kenett, R. S., ... & Agmon-Levin, N. (2022). A clinical pathway for the diagnosis of sesame allergy in children. *World Allergy Organization Journal*, 15(11), 100713.

Pearl, J. (1985). Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning" (UCLA Technical Report CSD-850017). Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, 329–334.

Pearl, J. (2015). Generalizing experimental findings. *Journal of Causal Inference*, 3(2), 259-266.

Shahin, M., Chen, F. F., & Hosseinzadeh, A. (2024). Harnessing customized AI to create voice of customer via GPT3. 5. *Advanced Engineering Informatics*, 61, 102462.

Shmueli, G. (2010) To Explain or to Predict? *Statistical Science*, 25(3), 289–310.

Weiss, A., Beloosesky, Y., Kenett, R. S., & Grossman, E. (2016). Change in systolic blood pressure during stroke, functional status, and long-term mortality in an elderly population. *American Journal of Hypertension*, *29*(4), 432-438.

Yahia, S. H., Machnes-Maayan, D., Frizinsky, S., Maoz-Segal, R., Offenganden, I., Kenett, R. S., ... & Kidon, M. I. (2022). Oral immunotherapy for children with a high-threshold peanut allergy. *Annals of Allergy, Asthma & Immunology*, *129*(3), 347-353.

Zinger, G., Kepes, N., Kenett, R., Peyser, A., & Sharon-Gabbay, R. (2023). A Multivariate Meta-Analysis for Optimizing Cell Counts When Using the Mechanical Processing of Lipoaspirate for Regenerative Applications. *Pharmaceutics*, 15(12), 2737.

neaman.org.il

Statistics

Samuel Neaman Institute for National Policy Research
Technion City, Haifa | +972-4-8292329 | info@neaman.org.il