The Importance of Probability and Statistics in Engineering and Computer Science

Sheila Gobes-Ryan, Ph.D., University of South Florida

January 15, 2023

Increasing Role

You are entering your professional lives at a time when how we do business is radically changing because of technology and the data collection and analysis that it enables. The Global Datasphere: all the data created, captured, or replicated; is predicted to grow to 175 zettabytes (ZB) by 2025 (Reinsel, Gantz & Rydning, 2018). (A zettabyte equals $1{,}000^7$ bytes (Desjardins, 2019).) To put this growth in perspective, the Global Datasphere was 33 zettabytes in 2018 (Reinsel, Gantz & Rydning, 2018). To enhance your professional success, you will need “the ability to read, work with, analyze, and argue with data as part of a larger inquiry process” (D’Ignazio & Bhargava, 2016, p. 84). Therefore, learning to make data professionally valuable is a large part of any probability and statistics course.

The words we use - Connecting Probability and Statistics with data and data literacy

To use probability and statistics professionally, you must find reliable data, apply that data to address projects, and effectively present what you have done to various audiences. Professionals who can work with data in this way are ‘data literate,’ a skill that is becoming increasingly important in engineering, computer science, and many other fields.

Where are we getting data?

Data is routinely collected in many facets of daily life by the systems with which we regularly interface. Therefore, it is professionally vital to be able to evaluate data credibility and identify data sources. Data may be available internally through your organization or your clients. It may also be purchased or available through external organizations or government entities. For example, US government data is collected and made available through many agencies, including NASA, the Environmental Protection Agency, the National Geologic Survey, the Bureau of Labor Statistics, and the Department of Transportation. Categories in which data relevant to your fields are available include national defense, foreign aid, energy and environment, trade, transportation, health, and climate. You can start with the USA Facts website for basic data visualizations with links to more detailed data sets. In addition, many countries, regions, and international governmental organizations have similar resources, including Eurostat, Statistics Bureau of Japan, and OECD data.

Other sources

Increasing numbers of non-governmental organizations assemble and make data available for use. A few examples are Data World, Gapminder, and Climate Trace. (Gapminder also makes available excellent data visualization tools.) Beyond these organizations are businesses specializing in collecting, assembling, and selling data.

Phases of Work

Problem Definition

When a client comes to you with a problem or project, developing a suitable solution involves ensuring you understand the specific issue clients hire you to address. Obtaining data surrounding the issue can go a long way to making the specifics of what is currently happening clear and offer means to predict future trends.

Examples

Engineers *determine signal control settings and road and bridge expansion* with statistical and probabilistic traffic data analysis (Abdel-Rahim, n.d., Mugdha, n.d.).

Prediction of workforce demand is vital to keep many of our daily systems working smoothly. Several areas include supply chain planning, hospital staffing, retail staffing, and TSA staffing at the airport (Sefair, 2021).

Health tracking devices collect data used to understand health issues in more detail. One example is the 'Smart Bra' to collect data to predict the indications of heart problems in women (Weintraub, 2018).

Design Phase Evaluation

As Engineers and computer scientists develop products and solutions, they must test the performance of these new products and solutions. For some products, this can involve understanding customer needs and preferences through user testing and analyzing it statistically. For others, developers test new products or systems to evaluate how they perform within a larger population. Finally, we can assemble and analyze numerous data sets to predict product performance across multiple characteristics.

Examples

Conjoint analysis of customer preferences: In this process, professionals design an experiment to test several modifiable product characteristics. The designers identify potential customers to test a product and respond to different product characteristics. Hahn and Doganaksoy (2008) discuss this approach being used to evaluate the characteristics of a new bar soap, including weight, size, shape, color, surface feel, type of fragrance, and fragrance intensity.

Energy consumption models for buildings: These models use multiple data inputs from past conditions influencing a building's energy use to predict the performance of a building from a digital model (Catalina, Virgone, & Iordache, 2011). These data include wind direction, wind speed, temperature, and orientation to the sun.

Process Tracking and Prediction

Devices collect data on their operation and on the people and processes using them. The volume and scope of data collected allow for specific and detailed analyses that were not possible a generation ago. These data are used for predictive maintenance, monitoring, and analytics, enabling increased efficiency and process optimization.

Example

Nuclear power plant safety, security, and safeguards. The need for security at nuclear power facilities and the data generated from their operational processes afford unique challenges balancing security, risk, and complex processes. To address these challenges, government and research organizations are developing statistical methods so that only data needed for operational and predictive maintenance are accessible (Sandia National Laboratories, 2018).

Product Performance

Manufacturers assess product performance is assessed to ensure that a products meet performance expectations set by the manufacturer, customers, or standards organizations (i.e., ASTM, NFPA, ISO).

> ***Example***
>
> *Software reliability growth models.* These inferential statistical tools are used in the early testing of software to describe an application's failure in terms of its internal structure to assess its reliability (Febrero, Calero, & Moraga, 2014).

Business Evaluation

Collection and evaluation of statistical data are foundational for businesses and their leadership. Data allows organizations to evaluate market demand, details of business performance, and identify problems, as well as to provide predictive analytics for decision-making on issues such as business location (Richards, n.d.a, Richards, n.d.b)

> ***Example***
>
> *Profitability.* In the article "The True Measure of Success," Michael Mauboussin (2012) describes the importance of knowing profitability by customer. He discusses how he discovered that his organization's largest customers were the most time consuming and least profitable. In contrast, their mid-sized customers were less demanding and more profitable. This analysis resulted in a move to the more profitable client type.

Software & Data Visualizations

Probability is essential for creating algorithms that allow two possibilities in software. The first is that probabilistic algorithms make random choices possible in computing. The second, the probabilistic analysis of algorithms, incorporates randomness into the data processed by an algorithm (National Research Council, 1992).

Computer scientists use statistics and probability to create data visualizations essential for representing data in straightforward and understandable ways. This approach, when used well, tells stories with data that would not be clear or accessible to broad audiences in other formats. In addition, probability allows businesses to use data to reduce future uncertainty through scenario analysis, sales forecasting, risk management, and budget forecasting.

Examples

Communication protocols introduce randomness for multiple packets of information using the same communication channels, allowing them to avoid being in the same place simultaneously in the communication system (National Research Council, 1992).

Pandemic data. Johns Hopkins created the widely used COVID-19 dashboard to visualize current COVID-19 data in several ways globally and by country (Johns Hopkins Coronavirus Resource Center, 2022) This presentation allowed many people to easily access and understand massive amounts of data on COVID-19 numbers and distributions.

Conclusion

The examples above are a few of the many uses of probability and statistics in Engineering and Computer Science. With the quantity of data growing year by year, it is essential to identify the types of problems and questions that statistics and probability can help you address so that you are able to successfully use these tools as part of your professional practice.

References

Abdel-Rahim, A. (n.d.) Models of traffic flow. University of Idaho. https://www.webpages.uidaho.edu/ahmed/ce322/class%20notes/class%2015%20_traffic_stream_parameters.pdf

Catalina, T., Virgone, J., & Iordache, V. (2011, November 14-16). Study on the impact of the building form on energy consumption [Conference paper]. *Proceedings of Building Simulation 2011*. pp. 1726-1729. 12th Conference of International Building Performance Simulation Association, Sydney, Australia.

Desjardins, J. (2019, April 17). How much data is generated each day? World Economic Forum. https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

D’Ignazio, C., & Bhargava, R. (2016). DataBasic: design principles, tools and activities for data literacy learners. *The Journal of Community Informatics 12*(3) 83-107.

Febrero, F., Calero, C., & Maraga, M. Á. (2014). A systematic mapping study of software reliability modeling. *Information and Software Technology 35* 839-849. Doi: 10.1016/j.infsof.2014.03.006

Hahn, G. J., & Doganaksoy, N. (2008). *The role of statistics in business and industry*. [ebook] John Wiley & Sons. http://ebookcentral.proquest.com/lib/usf/detail.action?docID=819142

Johns Hopkins Coronavirus Resource Center (2022). *COVID-19 Dashboard* [Data Visualization]. https://coronavirus.jhu.edu/map.html

Mauboussin, M. J. (2012, October) The true measure of success. *Harvard Business Review. https://hbr.org/2012/10/the-true-measures-of-success*

National Research Council (1992). Probability and Algorithms. Washington, DC: The National Academies Press. https://doi.org/10.17226/2026.

Reinsel, D., Gantz, J., & Rydning, J. (2018, November). The digitization of the world: From edge to core [White Paper]. International Data Corporation. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Richards, L. (n.d.a). *The role of probability distribution in business management.* The Houston Chronical Chron website, Small Business. https://smallbusiness.chron.com/role-probability-distribution-business-management-26268.html

Richards, L. (n.d.b) *How confidence interval affects business.* The Houston Chronical Chron website, Small Business. https://smallbusiness.chron.com/role-probability-distribution-business-management-26268.html

Sandia National Laboratories. (2018). *Industrial internet-of-things & data analytics for nuclear power & safeguards.* https://www.osti.gov/servlets/purl/1481947

Sefair, J. A. (2021, December 6). Dynamic workforce3 allocation to improve airport screening operations [Faculty Candidate Presentation]. University of South Florida, Industrial and Management Systems Engineering

Weintraub, K. (2018, June 27). This bra could save lives: An MIT startup reimagines the bra as a heart monitor--and aims to fill the data gap for women's heart disease. *The MIT Technology Review.* https://www.technologyreview.com/2018/06/27/240494/this-bra-could-save-lives/